CATALYST: A CLOUD-BASED DATA CATALOG SYSTEM FOR A SWEDISH MINING COMPANY

Adyasha Swain

Computer Science and Engineering, master's level (120 credits)
2019

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering
Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering
Master Programme in Computer Science and Engineering-
specialization in Distributed Cloud Computing

Adyasha Swain

CATALYST: A CLOUD-BASED DATA CATALOG SYSTEM
FOR A SWEDISH MINING COMPANY

Examiner: Chaired Professor Christer Åhlund (Luleå University of Technology)

Supervisors: Associate Professor Saguna Saguna (Luleå University of Technology)
Associate Professor Karan Mitra (Luleå University of Technology)
Markus Frank (Boliden)
ABSTRACT

Luleå University of Technology
Department of Computer Science, Electrical and Space Engineering
Master Programme in Computer Science and Engineering-
specialization in Distributed Cloud Systems

Adyasha Swain

CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company

Master’s Thesis

96 pages, 58 figures, 17 tables

Keywords: Internet of Things, Cloud Computing, Big Data, Data Catalog, Data Isolation.

In today’s digitization scenario, drivers such as the Internet of Things (IoT), cloud computing and big data lead to many initiatives such as Industry 4.0 or smart manufacturing. Large mining organizations are witnessing the emergence of big data not only through IoT but also through legacy systems and internal processes. Addressing big data is a challenging and time-demanding task that requires an extensive computational infrastructure to ensure successful data processing and analysis. Though most organizations have adopted a wide variety of powerful analytics, visualization tools, and storage options, efficient data usage, and sharing is taxing and may lead to data isolation. The thesis proposes, develops and validates a data catalog system called CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company to address the data isolation, access and sharing challenges faced in a large organization. The prototype implementation and the evaluation of our system show that the average query time was reduced from 59.813 milliseconds to 11.009 milliseconds, as well as the average data count was reduced from 12,691 to 5721.7,
which is almost less than 50 per cent, and solving data isolation challenges within Boliden, a large Swedish mining company. Finally, Boliden has confirmed the value of CATALYST in general and finds it beneficial for data management within their organization.
ACKNOWLEDGEMENT

The last two years of my life has given me a lot of memories to cherish. It is through this program that I was able to experience different culture by making good friends. In these two years, I had my own hard times where deadlines for assignments, presentations were exhausting. Fortunately, through those time until now, I have a great support from people when it was needed the most.

I want to express my sincere gratitude to my supervisors for being very supportive from the beginning. Thank you for your patience, motivating words which helped me to grow as an individual. Professor Saguna and Professor Karan Mitra constantly advised me on how to improve my writing skills for reports, doing better presentation for assignments, ways to project my ideas for project courses.

I am very grateful to Frank Markus for giving me the opportunity to work under Boliden for my thesis project. Thank you for always listening and constantly guiding me throughout this entire period. I always appreciated your approach and insight to my work, which helped me to improve.

Special thanks to my parents and grandparents for constant support. They would always calmn me down by their encouraging words whenever some situation stressed me out. I dedicate this thesis to you.

Adyasha Swain
# TABLE OF CONTENTS

1 Introduction .................................................................................................................. 10
   1.1 Background ............................................................................................................. 10
   1.2 Research Motivation ............................................................................................. 13
   1.3 Motivation Scenario .............................................................................................. 14
      1.3.1 Challenges and its implications ...................................................................... 16
   1.4 Research Questions and Objectives ....................................................................... 18
   1.5 Research Methodology ......................................................................................... 18
   1.6 Thesis Goal ............................................................................................................ 19
   1.7 Delimitations .......................................................................................................... 20
   1.8 Thesis Contribution ............................................................................................... 20
   1.9 Thesis Outline ........................................................................................................ 21

2 Background and Related Work ....................................................................................... 22
   2.1 Introduction ............................................................................................................. 22
      2.1.1 Cataloging Needs ............................................................................................ 23
   2.2 Data Cataloging ...................................................................................................... 25
   2.3 Enabling Technologies ............................................................................................ 27
      2.3.1 IoT .................................................................................................................. 27
      2.3.2 Analytics ......................................................................................................... 29
      2.3.3 Cloud Computing ........................................................................................... 30
      2.3.4 Big Data .......................................................................................................... 33
      2.3.5 Data Lake ........................................................................................................ 35
      2.3.6 Data Warehouse ............................................................................................. 36
   2.4 Related Work .......................................................................................................... 37
      2.4.1 Need for Cataloging in Commercial Applications and Industry ................. 37
      2.4.2 Current Research in Cataloging ........................................................................ 40
   2.5 Discussion .............................................................................................................. 43
   2.6 Summary ................................................................................................................ 44

3 CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company ......................................................................................................................... 45
   3.1 Introduction ............................................................................................................. 45
   3.2 System Overview .................................................................................................... 47
LIST OF FIGURES

1. Gartner’s 2014 Hype cycle. .................................................. 10
2. Data points and audiences. ..................................................... 11
3. Scenario of a data-driven organization. ................................. 12
4. Different dimensions of data which needs to be understood. .... 13
5. Challenges in using data and analytics capabilities[86]. ........... 15
6. Desired Boliden Data and Analytics Framework ....................... 16
7. Boliden’s Current BI/BA Situation. In the background is the Boliden business information service component, also shown in figure 6. 17
9. Components need to assimilate together for provision of placid response. ......................................................... 22
10. Transformation of integration towards application specific integration flows. .............................................................. 24
11. Gartner’s inquiry on data catalog, 2017. .............................. 25
12. Overall generalised view of the approach used in this thesis. ....... 26
13. IoT technology roadmap[35]. .................................................. 28
14. Different levels of analytics process. ........................................ 30
15. Compound Annual Growth Rates (CAGR) by Cloud Services Category[16]. ................................................................. 32
16. Big data research across the disciplines of publications supported by NSF ................................................................. 33
17. Operational data lake. ............................................................ 35
18. Typical data warehouse architecture. ...................................... 36
19. Modern data preparation pipeline[85]. ................................... 42
20. Leading vendors in the market[21]. ........................................ 43
21. Facts about Boliden. ............................................................. 44
22. Numerous options to store and process the data in Microsoft Azure platform[74]. ................................................................. 46
23. Service offerings by Microsoft Azure cloud platform. ............... 47
24. CATALYST Architecture layer view ....................................... 48
25. Organization dealing with structured data. ............................ 52
26. Organization dealing with structured data in a multi parallel environment. ................................................................. 53
27. Organization dealing with unstructured data. ......................... 54
28. Organization dealing with structured data and unstructured data. ................................................................. 55
29. Data Governance requirement in an enterprise. ...................... 56
<table>
<thead>
<tr>
<th>Page</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>30</td>
<td>Data Governance’s maturity sketch.</td>
</tr>
<tr>
<td>31</td>
<td>Comparison between the three stages of data governance approach.</td>
</tr>
<tr>
<td>32</td>
<td>Differentiation between centralized and agile data governance.</td>
</tr>
<tr>
<td>33</td>
<td>GSACTIVITYDATAKINDVALUES: Excel file machine activity data details from Kristineberg.</td>
</tr>
<tr>
<td>34</td>
<td>GSACTIVITYDATATLOG: Excel file machine activity data details from Garpenberg.</td>
</tr>
<tr>
<td>35</td>
<td>GSPUBLISHEDBLASTS: Excel file blast excavation data details from Garpenberg.</td>
</tr>
<tr>
<td>36</td>
<td>Data catalog database dashboard.</td>
</tr>
<tr>
<td>37</td>
<td>Registration for a data source into data catalog.</td>
</tr>
<tr>
<td>38</td>
<td>Published gannt data to data catalog, page 1.</td>
</tr>
<tr>
<td>39</td>
<td>Published gannt data to data catalog, page 7.</td>
</tr>
<tr>
<td>40</td>
<td>Published Garpenberg budget report into data catalog.</td>
</tr>
<tr>
<td>41</td>
<td>Preview of the budget report in the data catalog.</td>
</tr>
<tr>
<td>42</td>
<td>Power BI implementation procedures for reporting.</td>
</tr>
<tr>
<td>43</td>
<td>Architecture for the implementation.</td>
</tr>
<tr>
<td>44</td>
<td>SQL database for the implementation.</td>
</tr>
<tr>
<td>45</td>
<td>Code snippet for catalog evaluation.</td>
</tr>
<tr>
<td>46</td>
<td>Request send for the production data.</td>
</tr>
<tr>
<td>47</td>
<td>Query and subqueries generated for the request made.</td>
</tr>
<tr>
<td>48</td>
<td>Code snippet for non-catalog evaluation.</td>
</tr>
<tr>
<td>49</td>
<td>Request send for data in non-catalog scenario.</td>
</tr>
<tr>
<td>50</td>
<td>Query and subqueries generated for non-catalog scenario.</td>
</tr>
<tr>
<td>51</td>
<td>Comparison of the same query made for gannt data.</td>
</tr>
<tr>
<td>52</td>
<td>Comparison of the same query made for gannt report.</td>
</tr>
<tr>
<td>53</td>
<td>Comparison of the same query made for budget report.</td>
</tr>
<tr>
<td>54</td>
<td>Outcome of both the perspective.</td>
</tr>
<tr>
<td>55</td>
<td>Comparison of query time between catalog and non-catalog approach.</td>
</tr>
<tr>
<td>56</td>
<td>Comparison of data counts between catalog and non-catalog approach.</td>
</tr>
<tr>
<td>57</td>
<td>Overall comparison of data counts between catalog and non-catalog approach.</td>
</tr>
<tr>
<td>58</td>
<td>Recommendations for tagging in the catalog for Boliden.</td>
</tr>
</tbody>
</table>
# LIST OF TABLES

## List of Tables

<table>
<thead>
<tr>
<th>Table No</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Data and analytics challenge and its implications.</td>
<td>17</td>
</tr>
<tr>
<td>2</td>
<td>Interoperability on various levels.</td>
<td>23</td>
</tr>
<tr>
<td>3</td>
<td>Participants of the meeting.</td>
<td>60</td>
</tr>
<tr>
<td>4</td>
<td>Annotations and its definition.</td>
<td>65</td>
</tr>
<tr>
<td>5</td>
<td>Type - Resource group</td>
<td>66</td>
</tr>
<tr>
<td>6</td>
<td>Type - SQL server</td>
<td>66</td>
</tr>
<tr>
<td>7</td>
<td>Type - SQL database</td>
<td>66</td>
</tr>
<tr>
<td>8</td>
<td>Type - Data Catalog service</td>
<td>66</td>
</tr>
<tr>
<td>9</td>
<td>Delimitated data points for the implementation.</td>
<td>73</td>
</tr>
<tr>
<td>10</td>
<td>Individual query hit details for gannt data.</td>
<td>75</td>
</tr>
<tr>
<td>11</td>
<td>Individual query hit details for gannt report.</td>
<td>75</td>
</tr>
<tr>
<td>12</td>
<td>Individual query hit details for budget report.</td>
<td>76</td>
</tr>
<tr>
<td>13</td>
<td>Individual query hit details for gannt data.</td>
<td>78</td>
</tr>
<tr>
<td>14</td>
<td>Individual query hit details for budget report.</td>
<td>78</td>
</tr>
<tr>
<td>15</td>
<td>Individual query hit details for gannt report.</td>
<td>79</td>
</tr>
<tr>
<td>16</td>
<td>Aggregated query time.</td>
<td>81</td>
</tr>
<tr>
<td>17</td>
<td>Aggregated data counts</td>
<td>82</td>
</tr>
</tbody>
</table>
LIST OF SYMBOLS AND ABBREVIATIONS

Acronyms

BA  Business Analytics.

BI  Business Intelligence.

CPS  Cyber-Physical System.

ERP  Enterprise Resource Planning.

IAAS  Infrastructure as a Service.

IoT  Internet of Things.

MPP  Multi Parallel Processing.

PAAS  Platform as a Service.

SAAS  Software as a Service.

SLA  Service Level Agreement.
1 Introduction

This chapter discusses the context in which this thesis takes place, explaining the background which identify the scenario related to the bigger overview about the data driven organization.

1.1 Background

It is no secret that digitization holds great promise for all industry, lifestyles, societies and can fundamentally transform every aspect of information to be accessible and shareable. Targeting manufacturers by tradition; mining and manufacturing has been thought to be a process that turns raw materials into physical products. But in reality there are fragmented communications protocols, automation practices, supplying vendors, stakeholders, and many more entities that goes into one supply chain which has to be managed\[48\][7]. In this vision, billions of sensors, actuators, intelligent machines and everyday objects are expected be connected to the Internet and communicating with each other without any human intervention; and also solutions for the storage of these data generated in the process. This coined terminologies like Internet of Things and Cloud computing; and has become one of the hyped technologies regarding business and technology context, as shown in figure 1.

![Figure 1: Gartner’s 2014 Hype cycle.](image-url)
According to Gartner report\[78\], it is expected to see 20 billion internet-connected things by 2020. As said by Mark Hung, Gartner Research Vice President “Initially, leaders viewed the IoT as a silver bullet, a technology that can solve the myriad IT and business problems that their organizations faced. Very quickly, though, they recognized that without the proper framing of the problems, the IoT was essentially a solution looking for a problem” \[34\]. It is verifiable that, it will have a great impact on the economy by transforming many enterprises into digital business and facilitating new models, improving efficiency and increasing customer engagement [bloom\textsuperscript{2019}]. However, the ways in which enterprises can actualize and benefits will be diverse and, in some cases complex.

Figure 2: Data points and audiences.

The term Industry 4.0 was manifested for the first time at the Hannover Fair with the presentation of the “Industry 4.0” initiative; which marked the dawn of the fourth industrial revolution. The concept of industrial 4.0 can be characterized as a transformation of production into a fully automated and optimized manufacturing environments; where sensors, machines and IT systems are interconnected within the value chain across the enterprise boundaries. It leads to the integration of the concept of Cyber-Physical System (CPS) with Internet of Things (IoT); which uses the cloud-based manufacturing for
creating, publishing, and sharing the services that represent manufacturing processes. This becomes the cornerstone for smart factories\textsuperscript{19}; where requirements are to built a system which are highly adaptable and can utilize resources efficiently; combining all the physical and software components to perform a certain task. It leads to the trend of having both information and services at hand; making data exchange as one of its important feature.\textsuperscript{46}. At present, the majority of the manufacturing plants and production facilities around the world are planning to place such systems\textsuperscript{68} \textsuperscript{73}. It will increase global competences which will require an integration of systems across domains by forcing the definition of new road-maps for the enhancement of processing power \textsuperscript{12}.

This brings big data challenges; as most enterprises do not know what to do when an organization have huge amount of data points with varied structure, potentially millions of data assets and thousands of people consuming it, see figure 3. Despite the fact that most organization have adopted a wide variety of powerful analytics and visualization tools\textsuperscript{40} \textsuperscript{60}; sharing data knowledge through those tools is quite grueling; arousing questions like how will humans and machine communicate with each other? How do you turn plenty of data at your fingertips into a meaningful insight? How do you find the right data to fit your analysis?\textsuperscript{70}.

As figure 4 suggests there should be one place for anyone within the organization to find curated data, understand how that data has been used? What is the life cycle of the data? And why it was created? And trust that it is right for the analysis at hand, whether

\begin{figure}[h]
\centering
\includegraphics[width=\textwidth]{figure3.png}
\caption{Scenario of a data-driven organization.}
\end{figure}
they are a data scientist, an analyst, or even a casual business consumer of data. There should be a practice of shareable knowledge rather than tribal knowledge, where unwritten information about a data is not shared among participants in an organization. We will be using this terminologies later on in our work, for better understanding of data management.

Figure 4: Different dimensions of data which needs to be understood.

In this context, we consider the contrasting data points with unstructured, semi-structured and structured data sets; and deriving value out of it by comprehending the approach for a data catalog and better reusability of data sources for further processing. We evaluate the approach based on the exploration made by receiving assessment from different departments in and across an organization.

1.2 Research Motivation

In the press release article by McKinsey on “The age of analytics competing in a data driven world”, proves that in today’s data-driven world as organization deal with large and fast-growing sources of data or information; also present a complex range of analysis and use problems. As the big data’s potential keeps growing; the inevitability of a data-driven society powered by big data analytics has long been a common practice within
and across industries. But while doing so; it brings vulnerabilities. While the heterogeneous data volume grows exponentially; the ability to analyze the data becomes complex. It is required to unlock the value of the data by identifying ambiguous patterns and associations; which can be resolved through advanced analytics that turns data into information.

But with advancement in technologies and algorithm, the gap to unlock the value of the data widens. It prevents enhancement in the supply chain of an organization; as it will lead to data isolation. Gartner defines it as the information assets organizations collect, process and store during regular business activities, but generally fail to use for other purposes such as analytics, reporting, business relationships and direct monetizing. Within the digital world, it is characterized as “hidden” or “uncategorized, unmanaged, and unanalyzed”. By late 2017, it was estimated that 80 percent of existing data falls under this category; with that figure expected to reach 93 percent or higher by 2020. At the same time, an exploding “internet of things”, integrate up to 212 billion data collecting devices by 2020. Organizations are collating and storing more data, most of it unstructured, than other time in human history. As data has become an important asset; its value can be enhanced; if it can be reused for different purposes; which may not be known at the time of collection.

As we can see in the figure 5 that the challenges in the data and analytics domain are huge and acquiring with effective implementing systems that use these technologies is neither easy nor cheap and can impose vulnerabilities which can be risky. The challenges with data isolation could be lessen; if the contents of existing enterprise data assets remain transparent to its consumer and producer. The concept of data catalog in a broader scene seems to fit in, as it maintains an inventory of data assets through the discovery, description, and organization of datasets. In the extreme, this can also be seen as a lifeline in a decentralized model to enable cross-functional data to use and slowly build common data analytics capabilities.

1.3 Motivation Scenario

Boliden AB is a Swedish mining and smelting company focusing on production of copper, zinc, lead, gold and silver. Its overall strategy is to create profitable growth adapting to recent trends in the market and investments in competitiveness and organic growth; and
by evaluating opportunities for acquisitions. For Boliden data derived insight and decision are getting increasingly important, as shown in the figure 6, is the desired framework of its own organization.

Though many sites within Boliden have in pockets a high maturity, but when it comes to analytics capabilities the overall maturity for Boliden is low. The figure 7 states the current situation of the organization.

A strategy has been set and several projects are ongoing such as reporting and visualization of processes data, setting up of a production data platform for mines, ideas on a common data platform for smelters as well as activities coordinate by the automation’s programs. Historically, Boliden is operating in a decentralized model, but in reporting and analytics field, the business areas and sites have started loosely coordinated activities. It would not be wrong to mention that it has huge roadblocks ahead but could be surpass if proper initiatives are taken. IT and Mines Technology department have taken an increased ownership in the area and have set up a global data analytics service team.
1.3.1 Challenges and its implications

The table 1 below states the possible data and analytics challenges and its implications. The key challenge regarding to thesis is that for a considerable time forward there will be loosely coordinate activities in data analytics area with a limited opportunity to standardize existing solutions. This means information might not be spread, insight and data are only available in pockets and key potentials for productivity might be missed. This also mean potentially duplicated effort and a heavy focus on non-value activities in data and analytics projects.
Figure 7: Boliden’s Current BI/BA Situation. In the background is the Boliden business information service component, also shown in figure 6.

<table>
<thead>
<tr>
<th>Challenges</th>
<th>Implications</th>
</tr>
</thead>
<tbody>
<tr>
<td>Diverse application landscape with number of different sources</td>
<td>Long process to build up common capabilities</td>
</tr>
<tr>
<td>High use of Excel data sources and manual data capturing processes, e.g. for production numbers</td>
<td>Data quality and data capture can be an issue</td>
</tr>
<tr>
<td>Sensor data often needs further processing and understanding before it can be used in reporting, e.g. a 0 or value missing can mean different things</td>
<td>Data quality can be an issue and high initial effort needed</td>
</tr>
<tr>
<td>Many activities are done with external partners due to internal bandwidth and competence</td>
<td>Individual setup of data and analytics activities</td>
</tr>
<tr>
<td>Reporting and visualization activities in sites have been set up differently per use case</td>
<td>Missed synergies and high maintenance costs</td>
</tr>
<tr>
<td>Focus on recent activities is on delivery, not in building up data and analytics capabilities</td>
<td>Development of data and analytics capability is slower than it could be</td>
</tr>
<tr>
<td>Common master data management exists to some extent on Mining side but not overall, and needs governance and stewardship</td>
<td>Risk for data quality issues and rework</td>
</tr>
<tr>
<td>Missing guideline and standards on solution architecture, implementation approaches, and operation</td>
<td>A risk that solution diverse even more</td>
</tr>
</tbody>
</table>

Table 1: Data and analytics challenge and its implications.
1.4 Research Questions and Objectives

This section presents the questions and hypotheses that will be addressed and sets recommendations regarding next step for Boliden that should be achieved in this research as mentioned in our thesis goal.

1. Can we develop a system to efficiently gather data in large mining organization and making it available within the department?

The research challenge involves investigating the possible ways of publishing data centrally so that it does not get lost in the stockpiled of existing enterprise dataset. Also, investigating on how to provide context to a data to enhance the information sharing in a supply chain.

2. How efficient is our approach proposed by this research?

This research objective involves implementation our approach by presenting a real data set from Boliden as the use case scenario.

1.5 Research Methodolgy

Selecting the right research methodology and methods is a critical step for planning and performing the research work. Methodology defines the organization, designing, conducting and evaluating research. Moreover, appropriate selection of research strategy assures the quality of the conducted research. This section gives an overview of the most commonly used research methods and methodologies and justifies methodology selected for this work.

Figure 8 shows the system development research methodology used for our thesis work. It comprises of five steps: 1) construct a conceptual framework; 2) develop a system architecture; 3) analyze and design the system; 4) build a pilot system; and 5) observe and
evaluate the system. Meeting the goal of building a centralized catalog approach system to prevent data isolation depends on developing a conceptual framework to understand the characteristics of a system and its functionalities. The system architecture is important because it serves as a top-level structure that guides the development of the system. By examining relevant technologies, information systems researchers can adopt new approaches to analyze and design a more effective system.[39]

Figure 8: System Development Research Methodology Process Model[15]

1.6 Thesis Goal

The thesis goal is to understand the existing enterprise data management solutions which lead to an information silo, or a group of such silos. The data producers and data consumers are not aware of the priorities, responsibilities and the goals of other departments. This creates data isolation where there is little communication and collaboration within an organization. A system was built on top of a heterogeneous data source landscape which
will facilitate reuse of data for reporting and further processing with different targeting
group. In particular, it aims to comprehend whether the approach is well suited on Hor-
izontal and Vertical Integration of information based on a cloud infrastructure platform.
We aim to mimic the approach beyond the smaller-scale prototype to an enterprise service.
It will support in achieving the possible dimensions such as identifiable source, descrip-
tion of data, description of how to use, a reference to data, contact person, geo-location,
categorization of data and so forth.

1.7 Delimitations

Many initiatives for better decision making are undertaken in the competitiveness tech-
nology domain for generating better use cases and ease of use for the customers. Many
applications and services are used and well as outsourced in an organization to enhance
the analytical and reporting capabilities for a common need. Integrating them together
when voluminous varieties of data sources are consumed and produced at the same time;
is imposing great challenges to organize these data assets in a long run. We limit ourselves
to the investigation of our proposed system approach by using standard components as
much as possible; complaint with mining area of Boliden. Our main focus would be to
advocate; on how to setup data catalog in general which can handle different data assets
in and across organization. Particularly, integration into a holistic IT landscape between
different stages of production and the respective resource and the information flow within
a mine and across different sites along the value chain. The scenario that we would focus
will target data scientist and IT colleagues building reports in the mining industry and
recommendation for Boliden on next steps.

1.8 Thesis Contribution

The contribution to the thesis is the existing challenges regarding the congregation and
usage of existing enterprise data assets; which is growing multifold due to technology
advancement and services provided by vendors. These constraints where revealed here to
proof the relevance of the thesis work. The challenges in the industry market and research
field was analyzed to develop our approach and the following outcomes were achieved in
our work.

1. A centralized cloud-based shareable system naming “CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company” was proposed to manage the enterprise data assets in the big data and analytics field.

2. A comprehensive analysis of our proposal and recommendations were made for Boliden and limitations were observed.

The research questions were answered with comprehensive real data experiment, which provided valuable inputs to our proposal. This evaluation outcome can help Boliden to set up a shareable system using cloud based platform amidst the challenges faced to have a holistic approach in the data landscape.

1.9 Thesis Outline

We now present an outline of the content of the next chapters in this thesis.

Chapter 2 contains the investigation on the challenges faced by the organization and the literature review in the research field, which explains the technical context of work. We cover the enabling technologies and the needs to set up a central shareable system.

Chapter 3 illustrates the data landscape and the implementation approach for our thesis.

Chapter 4 elucidates the result obtained, and discussion about the outcome of our work based on the evaluation and the opinion of the people at Boliden.

Chapter 5 we finish by presenting our conclusions and recommendations for the future work.
2 Background and Related Work

2.1 Introduction

The previous chapter gave us an overview about organizations initiating on the journey towards data driven; where it has to deal with large and fast-growing sources of data. But with the technology advancement, it makes the vision of organizational data landscape broader. The implementation of data catalog concept requires appropriate technologies, which are mature in its own domain. It will support the enhancement of the supply chain by reducing the data isolation between the data source to its analysis. This will enable the information exchange, optimization and reusability of data throughout the whole process of operations[45]. But in real scenario; it is a complex process for applying this approach to the data and analytics platform of the future[20].

Figure 9: Components need to assimilate together for provision of placid response.

To provide a logical solution to an inquiry; it requires information flow to be implemented correctly; for achieving efficiency, data quality and individualization of each system. These information flow require tighter integration and association between various components and participants in an operation[27], as shown in the figure 9. It should be able to interact with other systems without application of special effort for integration. For example, from data ingestion to data storage; to its transformation; to its analysis and its visualization components should merge together for better response.
2.1.1 Cataloging Needs

In a supply chain each component has various aspects; from mechanical components to strategic objectives and business processes. For an analytics platform; interoperability must be established at various levels [77], as shown in the table 2.

<table>
<thead>
<tr>
<th>Physical level</th>
<th>Assembling and connecting manufacturing equipment or products</th>
</tr>
</thead>
<tbody>
<tr>
<td>IT level</td>
<td>Exchanging information or sharing services</td>
</tr>
<tr>
<td>Business level</td>
<td>Operation and objectives have to be aligned</td>
</tr>
</tbody>
</table>

Table 2: Interoperability on various levels.

When establishing interoperability in manufacturing environments, different dimensions of integration has to be assessed like [47].

1. Vertical integration: Information integration regarding sensor, control, production, manufacturing, execution, planning, etc. which, includes factory-internal integration from sensors and actuators within machines up to Enterprise Resource Planning (ERP) systems.

2. Horizontal integration: Integration into a holistic IT landscape between different stages of production and the respective resource and information flow within a supply chain, and across different sites along the value chain.

3. Integration towards engineering and the applications: It denotes information sharing rather than tribal knowledge, between product and service development and different organization environments. It benefits the establishment of production, when information about the products to be created should be available for planning and manufacturing configuration tasks, as well as during product development or the analysis behind its development, could be used for designing optimization. Though it is an ideal situation, it lacks holistic approach and not much has been looked into.
As mentioned in the figure 10, the traditional industrial value chain consisted of independently implemented systems, including hardware systems and software systems. It supports product design, production planning, production engineering, production execution and services. Each has its own data formats and models, not only making integration of them difficult but also to find discrepancy in analysis and reporting, if needed. Cataloging of these individuals’ will blur the boundaries between these systems and activities.

Rather than sequential and hierarchical system integration; there will be a network of connected things, processes and customers that will allow companies to interact with customers and suppliers much more rapidly, accurately and effectively and in a more secured manner. As a result, implementation of specific solutions and applications in an organization will not only focus on customer demands, but also on the application specific establishment of information access and workflows. The full adoption of data catalog for architecture principles to production environments could provide that support.
2.2 Data Cataloging

The market is evolving for tools in modern BI and analytics platform. It is because of their agility and ease of use; data preparation tools started out being used for self-service use cases by analysts and data stewards to accelerate the preparation of data for interactive analysis and data science\[17\]. It demands for catalog solutions; as organizations struggle to inventory distributed data assets to accelerate data monetization and also conform to regulations. The figure 11; shows the inquiries on data catalog has considerably increased in 2017 as a proof.

![Gartner Inquiries on "Data Catalog"](image)

Figure 11: Gartner’s inquiry on data catalog, 2017.

In Gartner’s paper on “Data Catalogs Are the New Black in Data Management and Analytics” it explains data catalog as “A data catalog maintains an inventory of data assets through the discovery, description and organization of datasets. The catalog provides context to enable data analysts, data scientists, data stewards and other data consumers to find and understand a relevant dataset for the purpose of extracting business value”\[86\]. Also refer to figure 12 for better understanding of the approach used in this thesis. With vast amount of data generation in an organization, this solution offers an ideal solution to this problem. It supports in creation of a metadata repository that leads data audiences to an enriched, inventory, shareable and collaborative platform with the data across the enterprise\[25\]. It enables data users such as data and analytics leaders to introduce data governance which is mentioned later in the chapter for managing data sprawl and information supply chain, in support of digital business initiatives.
The factors provided by it could have better implications for data and analytics leaders if data catalogs is exploited effectively by linking them to a broader enterprise information management needs as well as to analytics platform. It would enable organizations to address the challenges and realize longer-term benefits with a broader data catalog approach by collating and communicating the up-to-date information asset inventory that is available to the organization, providing data usage transparency with lineage and impact analysis, capturing metadata to enhance internal analysis of data use and reuse, query optimization and data certification, contextualizing information within its business usage by capturing, communicating and analyzing what data exists, where it comes from, what contexts it is used in, why it is needed, how it flows between processes and systems, who is accountable for it, what it means and what value it has. In effect, creating an information supply chain, to document and communicate the context, meaning and value of data that has been loaded to a data lake as data lakes store data in an untransformed raw form, including data of immediate interest, potential interest and data for which the
intended usage is not yet known.

2.3 Enabling Technologies

As mentioned previously in this section about the needs and data catalog approach; it specifies that some technologies have to be addressed in order to implement the concept. In actual scenario in an organization, the implementation requires a significant amount of time period after initiatives have been made on the prototypes build in the first phase. It is better to consider that the maturity of technologies in many cases does not correspond to the expectations placed on them. In this section some examples of enabling technologies are discussed.

2.3.1 IoT

The Internet of Things is a concept that was first developed by Kevin Ashton in 1999 in the context of supply chain management, where he described a system in which the physical world is connected to the Internet through ubiquitous sensors[22]. The concept that was polished over the years and can be defined as “a dynamic global network infrastructure with self-configuring capabilities based on standard and interoperable communication protocols where physical and virtual ”things” have identities, physical attributes, and virtual personalities and use intelligent interfaces, and are seamlessly integrated into the information network, often communicate data associated with users and their environments”[80]. As a market, the whole annual economic impact caused by the IoT could have a value of up to 2-6 trillion USD by 2025, making it accurate to say that it as one of the most important areas of future technology.

The true value of it in regard to enterprise can be fully perceived when the devices connected to each other would not only communicate with each other but also integrate with business intelligence applications, business analytics, inventory systems etc. Soon firms will invest in the IoT to redesign its factory workflows as adoption of this technology is gaining momentum through technological competitiveness pushing firms to innovate and transform themselves. According to figure 13, application area is just not stagnant to specific application but wide spread to different domains such as in Healthcare for
monitoring, preventive and rehabilitation scenarios, in smart hybrid energy grids to save energy and reduce emissions, in smart metering at substations for informing plant about the low, mid and high demands, in smart cities for creating benefits for citizens’ well being and also able to state the rules and policy for the city government and development. With this advancement brings many challenges at the enterprise level such as.

1. Data management where IoT sensors and devices are generating massive amounts of data that need to be processed and stored. The current architecture of the data center is not prepared to deal with the heterogeneous nature and sheer volume of personal and enterprise data. Consequently, organizations have to prioritize data and generate ideas based on the needs and its value generation.

2. As more data are available for processing and analysis, the use of data mining becomes a necessity as it possesses vulnerability issues where data consist not only of traditional discrete data, but also of streaming data generated from digital sensors in industrial equipment, automobiles, electrical meters, etc. These streaming data are about geo-location, movement, vibration, temperature, humidity, and even chemical changes in the air. Corrective processes has to be invoked to address immediate operational issues or for better strategic moves and also to circumvent chaotic scenarios where solutions can become complex and hard to achieve.
3. Although there is improvement in the productivity of companies and enhances the quality of data collated but brings privacy risk with it as a main factor. Lack of security and privacy will create resistance to adoption of the IoT by firms and individuals. It may be resolved by tagging properly and taking ownership of the data precisely and encouraging organizations to engage in data governance policies at each level of data production pipeline.

2.3.2 Analytics

It is a concept for business related analytics tools and techniques such as Business intelligence, Business analytics and Real-time processing. Its main focus is on the past and investigate the next elucidation. There is an immense use of algorithms, statistics and predictive models to procure inestimable insight from the data in form of, as stated below and in the figure 14.

1. Business Analytics (BA): Comprised of solutions used to build analysis models and simulations to create scenarios, understand realities and predict future states. The framework provides data access to a huge amount of data from different kind of sources. This is a so called self-service data preparation facility where the user only gets access to authorized data. This data may come from production systems, sensors from different machines, external sources or other sources. The user itself takes care of transforming the data into usable information and uses it in the end-user tools for creating reports, information analyse, data mining, etc.[42][79].

2. Business Intelligence (BI): It is a term that includes the applications, infrastructure and tools, and best practices that enable access to and analysis of information to better understand the business. Business intelligence aims to support decision making, and a BI system can also be called a decision support system. It stems out to the applicability areas such as business performance management, enterprise performance management, yield management, etc[79].
2.3.3 Cloud Computing

The National Institute of Standards and Technology of the United States of America (2010) provides the following widely cited definition of cloud computing.

"Cloud computing is a model for enabling ubiquitous, convenient, on-demand network access to a shared pool of configurable computing resources (e.g., networks, servers, storage, applications, and services) that can be rapidly provisioned and released with minimal management effort or service provider interaction. This cloud model is composed of five essential characteristics, three service models, and four deployment models."

The essential characteristics of cloud computing are:

1. On demand self-service – The audiences especially the consumers can demand resources without any interaction with the cloud service provider.
2. Broad network access – Cloud resources can be accessed by clients from any heterogeneous platform devices.

3. Resource pooling – Cloud resources such as storage, network, etc. can be accessed by consumers via virtualization without any knowledge of its location with multiple users.

4. Rapid elasticity – Acquisition and emancipation of resources according to the need giving an illusion of infinite capacity of resources.

5. Measured Service – Users are charged based on pay-as-you-go or pay-per-use model where resources can be monitored properly.

6. Performance – Cloud resource can be scaled depending on the application workloads.

The service models can be mainly characterized into three categories according to the distribution model of the offered resources.

1. Infrastructure as a Service (IAAS) – The consumer are actually provided storage or machines virtually where the user is able to deploy arbitrary software and services on top of the provisioned resources, with the provider manages the infrastructure.

2. Platform as a Service (PAAS)- The consumer is able to deploy its application by the support of the provider in terms of development tools, libraries, application programming interfaces, etc. with configuration settings. Consumer do no manage the underlying infrastructure.

3. Software as a Service (SAAS)- The consumer is provided with a complete application, that is accessible through diverse clients. This model is useful for users that do not need or want to manage an application and only care about getting a service.
Figure 15: Compound Annual Growth Rates (CAGR) by Cloud Services Category\[16].

The deployment model can be characterized into mainly four categories according to the type of organization and who intends to use it.

1. Public cloud is provisioned by a large industry group or any public organizations whether it be in academics, companies, etc.

2. Private cloud is provisioned exclusive to a particular single organization and its department accountable.

3. Hybrid cloud is a combination of multiple clouds which are bound by technologies to provide application and data portability.

4. Community cloud are for exclusive set of community formed with multiple organizations.

As figure 15 states the compound annual growth of cloud service usages, we get a glimpse of the resources provided by cloud vendors for cloud applications such as healthcare systems, energy systems, e-governance and many more applications which are soaring in the market. It leads to competitions in the market where vendors are striving to ideate new services to the consumers for the ease of functionality in their workplace. Possibility of having millions of interconnected things, deployment, resource provisioning, auto-scaling
of computational resources and advancement in the real time processing, etc. are some of the main research challenges for the cloud layer in an enterprise.

2.3.4 Big Data

The pervasive nature of today’s digital technologies and data-reliant applications have made the expression ”big data” widespread across other disciplines too such as sociology, medicine, biology, economics, management, information science to name a few. Figure 16 shows the research domain across various disciplines in relation to big data. Authors in 21 state the categorization of an information asset as follows: “It is the Information asset characterized by such a High Volume, Velocity and Variety to require specific Technology and Analytical Methods for its transformation into Value”[18].

![Figure 16: Big data research across the disciplines of publications supported by NSF](image)

In order to generate value from big data we need to understand the following important aspects:
1. Information as an asset for big data.

The quick expansion of big data which is still enlarging in its sphere, is because of its extensive degree with which data are created, shared and consumed in recent times. A noteworthy example at its nascent stage was the mass digitization of Google Books Library Project, which begun in 2004; the whole idea backing it up was to fully digitize more than 15 million printed books held in several university libraries, including Stanford, Harvard and Oxford[11]. The data-information-knowledge-wisdom hierarchy proposes a view on information which appears as data which are structured in a way to be beneficial and presents specific purpose or insight. Also, the scenario in which artificial objects, equipped with unique identifiers, interact with each other to achieve common goals, without any human interaction, goes under the name of Internet of Things, IoT[4], and represents a promising source of information in the age of Big Data.

2. Its characteristics can be described as[26]:

   a. Volume, meaning the vast quantity of data generation and its storage. Its size attributes to its value, insight and its consideration.

   b. Variety, meaning the category and its classification to which a data belongs. For example, structured, unstructured data which data information from images, text, audio, excel, etc.

   c. Velocity, meaning the momentum at which the data is generated and its consumption as processed entity. Precisely the frequency of generation and its handling, visualization and reporting.

   d. Veracity, meaning the quality of the data and its significance by proper analysis.

3. Technology as a necessity prerequisite.

   With the utilization of extensive amounts of data, technology comes to the forefront such as cluster analysis, genetic algorithms, natural language processing, machine learning, neural networks, predictive modelling, regression models, social network analysis, sentiment analysis, signal processing and data visualization for better understanding of a business value in an organization. Its application can be seen
in Healthcare, Information Technology, Manufacturing, Education, Media, etc. As every sphere of entity generates data; it pilots challenges, such as integration, validation, trustworthiness, security concerns, and many more which needs attention.

2.3.5 Data Lake

It is widely used in an organization to store massive amount of data where raw data can be stored with its purpose unknown, as shown in figure 17. The audiences are mainly the data scientists with high accessibility and easy to update. It can help nagging problems of data integration. Mike Lang, CEO of Reveltyix, a provider of data management tools for Hadoop, notes that “Business owners at the C level are saying, ‘Hey guys, look. It’s no longer inordinately expensive for us to store all of our data. I want all of you to make copies. OK, your systems are busy. Find the time, get an extract, and dump it in Hadoop’”. Previously approaches were based on predetermined schema which made it obligatory for all the user to follow a specific data model. Unlike this monolithic view of a single enterprise wide data model, the data lake relaxes standardization and defers modeling, resulting in a nearly unlimited potential for operational insight and data discovery and attaining transparency for the data by being as close to the data source and in raw format; putting an end to data silos.

Figure 17: Operational data lake.
2.3.6 Data Warehouse

Structured core corporate data that have been cleansed, organized and presented in a way that is understandable to the business is stored here. The data is always defined, mapped and modelled according to business rules. The users are basically the business professionals; which signifies that the data present in a warehouse has been used for a specific purpose within organization and the storage space is not wasted on data that was never used; manifesting for the data to be trustworthy and as close to visualization layer\[41\]. Figure 18 shows the typical architecture of data warehouse usage in an organization.

![Figure 18: Typical data warehouse architecture.](image-url)
2.4 Related Work

As the face of organization is changing towards huge volume of data to be analysed; which mostly relies on set of applications to communicate with, and provide services to today’s demanding consumer. It can be related to the collection, storing, and analyzing more granular information about product, people, transactions, and sensor-generated messages with high volumes to drive business processes. It is not just limited to an organization but ranges from large enterprises to government agencies. All of this data creates aggregation and analytic opportunities, using tools that leverage multicore server architectures. The challenge for the next decade will be finding ways to better analyze, monetize, and capitalize all of this information. It will be the age of big data and analytics for an organization initiating towards data driven goal; where the value of the heterogeneous data could be resolved by transforming it into an information quickly. In the following subsection, we see the perception about varied existing enterprise data assets, commercially in industries and in research field[76].

2.4.1 Need for Cataloging in Commercial Applications and Industry

In this segment we mention about different Institution with varied challenges they are facing on their road map for getting values out of their use case proposition.

1. United Parcel Service (UPS)

Is an American multinational package delivery and support chain management company. It will be precise to say that UPS is no stranger to big data since the company now tracks data on 16.3 million packages per day for 8.8 million customers, with an average of 39.5 million tracking requests from customers per day. It can be assumed that much of its acquired data comes from telematics sensors in vechles, for example, including speed, direction, braking, and drive train performance. The data is not only used to monitor daily performance but also assign UPS’s drivers’ route structure, which is an initiative called ORION (OnRoad Integrated Optimization and Navigation). It relies on online map data to configure drivers’ pickups and drop-offs in real time. The company is attempting to use data and analytics to optimize the efficiency of its 2000 aircraft flights per day[31].

Challenges faced by the UPS company
Juan Perez the Chief Information and Engineering Officer at UPS mentions “Big data at the organization is all about the business case, how effective are we as an IT team in defining a good business case, which includes how to improve our service to our customers, what is the return on investment and how will the use of data improve other aspects of the business”. According to him drilling down on a single data set in isolation and fail to consider what different data sets mean for other parts of the business is the an important scheme to look at. The re-use of data of information can have a significant impact. For example, about using delivery data to help understand what types of distribution solution work better in different geographical locations. Arousing questions on whether there be more access points? Should drivers take their own decisions depending on the situation of shipments? Which can be answered using new technologies, different data, and analytics. Earlier the dialogue used to be about buying technology, creating a data repository and discovering information. Now the conversation is changing about how to manage and make the data much more active and sharable across domain so higher level of collaboration can be attained and provides benefit for everyone for a better repository, less duplication and much more insight to a data without getting lost in the future.\[59\].

2. Caesars Entertainment

Formerly Harrah’s entertainment corporation has been one of the leading American gaming corporation in Nevada. Today Caesars is augmenting from traditional analytics capabilities with some big data technologies and skills, so as to respond in real time for customer marketing and service. Most of its data come from web clickstreams and from real time play in slot machines. One of main goal is to pay fanatical attention by automated means to ensuring that its most loyal customers don’t wait in lines, spotting service issues and experimenting with targeted real-time offers to mobile devices.\[14\].

Challenges faced by the Caesars Entertainment company

Rizwan Patel, IT director, commented,” When it comes to implementation, it is essential to select use cases that solve real business problems. That way you have the backing of the company to do what it takes to make sure the use case is successful.” To prioritize what data to include in the analysis and every possible source of data should flow properly across domain and to data lake where access of all the data that is relevant to the project. Effort should be invested on importing, cleaning and organize each data source so that expectations are maintained by precise insight to the each data being used for analysis for
a better use case.62

3. United Healthcare

United healthcare group Inc. is a health care company in Minnesota. Like many other large has been focused on structured data analysis for many years. Now, however, it is focusing its analytical attention on unstructured data like the data on customer attitudes that is sitting in recorded voice files from customer calls to call centers as level of customer satisfaction is increasingly important to health insurers, because consumers increasingly have choice about what health plans they belong to.28

Challenges faced by the United Healthcare company

According to Alex Barclay, Vice President of Advanced Analytics for UnitedHealthcare Payment Integrity “The idea is to use better judgement to ensure that once a claim is received we pay the correct amount, no more, no less, including preventing fraudulent claims. In doing so we have to identify mispaid claims in a systematic and consistent way, requiring to embrace broadening landscape.” To invest in the knowledge and data discovery to get more insights from the data to perform data enhancements including better structure, meta-data layers, new data sources, new applications tools and reporting49.

4. Airbnb

Airbnb deals with in global online marketplace and hospitality service via its websites. Like many startups, the number of employees at Airbnb has grown significantly over the past several years. In parallel there has been growth in both the amount of data and the number of internal data resources which brings in the dilemma of the prospective that though the growth of data resources is healthy but reflects the investments in data tooling to promote data-informed decision making.

Challenges faced by the Airbnb company

According to Elena Grewal, Director Head of Data Science “The challenges we are facing on effectively navigating a sea of data resources of varying quality, complexity, relevance, and trustworthiness. There is an intent to have a single lens into the data-space by providing as much as context as possible while observing tool access controls to the underlying data, which is nothing but transparency.” The view of the data landscape is merely one
of many. It is ensured that a data product is developed that provides universal value we talked to employees across departments, roles, tenure, and data literacy levels, to better understand their pain points and concerns around data. It was apparent that it was needed to develop a system that enabled a shift in thinking. Relying solely on tribal knowledge stifles data discovery. It is correct to say that the work creating a self-service data culture is not over yet[81].

In the press article release by MIT Sloan Management review where fortune 500 companies leaders where interviewed by Randy Bean[8], gives us a glimpse into how the captains of industry are thinking about big data and how their companies are changing because of new insights gleaned from big data analyses. It was envisioned that two data environments that coexist side by side. One is the traditional production operational environment, which has to be locked down. It’s hard to get things in there, it’s hard to get things out of there, but it’s stable. It is used for financial reporting, regulatory reporting, customer statements, those types of activity. This is not information that you want to change. But coexisting with that can be a "discovery" environment where analytics can be used to sift through new data and also to shift through traditional data, and this can be used to discern the new patterns that can later be incorporated within a production environment. So it is called "the new" and "the known." The new environment is focused on discovery, and the benefit that big data technology and processes bring is that it makes it possible to "load and go" – which means beginning to access and analyze all of your data without first going through the data engineering process, which is costly and time consuming. The benefit is that organizations can answer critical business questions in seconds rather than days, days rather than weeks, and weeks rather than months. It can be seen that companies are rushing to make good decisions by identifying better solutions and investments.

2.4.2 Current Research in Cataloging

It can be considered from the previous section that major decisions are being taken in this direction of a centralised solution. There is a high demand for outreaching the insight of a data asset, as more industries are facing digital disruption. Its consequences are leading to analytics, strategic inventories of data for reusability both internal and external to the
enterprise without IT assistance. Data and analytics leaders are struggling to respond to this urgent need due to their over-reliance on IT-centric tools for finding, cleaning and transforming relevant data, and making it accessible to the growing number of distributed users in the enterprise. Leading to worthless time being spend on data preparation than needed, barely little time for actual analysis. This results in data consumers and data producers respondents to take steps in actuating the performance of integration and preparation for better results in an enterprise.

Exploration are being done on how to expand the capabilities of modern tools for better recommendation in data preparation. The areas that are being explored are data exploration and profiling. This will enable users from different background in an organization to search, profile, inventory data assets, and tag or annotate data for future use case scenarios. It has to be a cumulative process of assimilating metadata based on the necessity use of data. This should not create confusion with the overall metadata management solutions that have a broad scope across the whole data and analytics program. For a user collaboration; it should facilitate the promotion of publishing, sharing, ownership with governance features such as to access a particular set of groups and shareability to what extent. These features are missing for self-service data preparation tools which exist in isolation. These introduces vulnerabilities such as business glossary terms, rules management, lineage and framework issues which has to be governed and audited properly by IT for its quality before promoting them.

As we see in the figure 19; mostly all the data preparation tasks start from Mode 2, where raw data are explored, transformed and improved. Along the line it converges and compliments Mode 1, where data preparation are done for an enterprise metadata management, such as to capture metadata, perform centralized analytics, identify and certify the sources. Activities in the modern data pipeline leads to business users; anchor multiple and frequently changing data sources. It brings challenges to its critical nature; for getting insight to a data. Doing tasks for facing these challenges requires investigation with new data sources and its varied types; where the organization is uncertain about the outcome success ratio and wanting to trial the hypothesis made without the need of IT support. It is actually here that the capabilities of the maturing tools are oppressing; allowing the organization to break down the barriers; of having to precisely understand the management of assets overall. It will be achieved once the Mode 2 experiment is effective.
and the business market wants to analyse its findings; ensuring that the pipeline has the capability in place to capture and share metadata from these data preparation flows bidirectionally. This process will literally ensure that whether the organization has the ability to even promote initiatives that started as “self-service” to governed data discovery, centralized analytics or any critical missions when required. It is important that modern tools should help the operationalization of data flows and build use cases to its refinement for better model. This brings the current market in a state of instability; where the range of vendor choices is intensifying trying to embed self-service data preparation with analytics and BI [21], making a few of them in the top run as shown in the figure 20. Recent years, significant sub markets have emerged in parallel to the main market offerings that represent a significant focus on either vision or execution, but do not address shareable data integration delivery requirements centrally when systems collaborate [67].
2.5 Discussion

Similarly, if the circumstances at Boliden are looked upon; it is not so different from other organizations. There are huge milestones to be undertaken with experiments and opting the better use cases prevalent to the situation. Boliden has a good number of sources and tools already today but need to take some basic steps in order to leverage data analytics; which implies more focal point on data and its life cycle in a data pipeline. In the figure 21 we can see the challenges being faced at Boliden currently regarding a holistic approach of establishing a common ground when large sensor and production data sets are being generated by the equipment at the mining site with over 23 BI and BA tools, 340 data source and 30 ongoing projects. It is an obvious fact that getting lost in this mammoth of information can be easy and if necessary steps are not initiated, the envision to leverage data analytics will be in vain and will demand investments which could be channelized in a better form. So, the cornerstone for Boliden is to present a sharable knowledge platform, where the transparency of data is maintained in each phase of the process from its assemblage to the value generated out of it.

Figure 20: Leading vendors in the market.
2.6 Summary

This chapter discussed the current challenges faced by the organization and in the research field for a holistic approach regarding enterprise data assets. Due to this, we identified a gap of the existing systems. There was a lack in the approach of centrally shareable system for data assets which prevented the understanding limited to tribal knowledge. However, the introduction of an approach that could address the fundamental scenario for establishing a common platform will lessen the burden of users consuming the data assets by manifold.
3 CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company

3.1 Introduction

From the previous chapter it was established that the newfound awareness for the potentially risk-laden data may reside within an organization and it cannot be ignored. It is a possibility of data mismanagement which need to be discontinued. It is believed that the data that lacks a present use initially will unlock hidden value within that data if complemented by analytical tools, storage services, and other services for example reporting services at respective stages in an organization. By respective stages, it means that the schema of the data from its collection stage to its visualization stage undergoes through different means of procedural practices. Adding to the complexity each stage is also handled by respective staff accountable for it in an organization. This will lead to tribal knowledge i.e. meaning set of unwritten rules and information known by just a group of individuals within an organization but not common to others. With the technology change and the demand challenge as mentioned in the previous chapter, it brings more additional responsibilities and challenges which necessitates the phase shift from tribal knowledge to shareable knowledge. The transformation would require multiplex agreements within the organization for each phase of the process starting from its collection to the stage when we create a value from it. This chapter presents a perspective of the complex processes complementing each other leading to a proposed architecture “CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company” for achieving a shareable system. The CATALYST architecture would facilitate better decision making in this data landscape of an organization, where the requirements are varied with numerous options to store and process the data as shown in figure 22.
Figure 22: Numerous options to store and process the data in Microsoft Azure platform.\[^{74}\]

As enterprise’s requirements are soaring high to achieve numerous goals in prediction, analyze and, BI and Reporting field with numerous offerings to integrate and manage it. It becomes difficult to combine all the scenarios and use cases when there are real time data or historian data to be processed. Microsoft Azure cloud computing platform was considered for our approach for its better service offerings as shown in the figure 23.
3.2 System Overview

Different organizations are dealing with one or more form of data from organizational systems, third-party vendors, web-logs, global positioning system (GPS), images, audio, video, devices, sensors and social media. Considerations must be given to the initial phase of its collection from different sources; to data storage and its processing in the intermediate phase; and reporting and analysis in the later phase. This lead to attain objectives such as integrated view from disparate data source, minimize silos and multiple versions of truth, centralize analytical data for many users to access, provide user friendly data structure which can evolve and adapt, and above all to realize the business value from the data. This section presents a detailed explanation of each layer and the common scenarios originating in an organization due to the data source layer, triggering the importance of cataloging. We perceive that each layer has its own distinguishable attributes accomplishing its own tasks of evaluation on a data, but also complementing each other to get a final insight of it. There can be possibilities of vulnerabilities for slightest omission or misinterpretation of glossary terms or other key considerations. This necessitates for proper data
governance in an enterprise which is explained later in this chapter.

Figure 24: CATALYST Architecture layer view.
3.2.1 Device Layer

This layer is as close to the data points from where the data originates. It can be termed as raw data whose usage and value it yet unknown. It is uncleaned and in varied format from different sources like smart applications, sensors, GPS, site images, laboratory experiment, machinery, etc.

3.2.2 Data Source Layer

This is the second layer as we move up the architecture hierarchy. Raw data that were generated from the sources are accumulated and stored in this layer. It involves data integration process where they undergo initial staging process and the other term for it is “landing zone”. As the name suggests it acts like an intermediate storage process with an approach of ETL(extract, transform, load) or ELT(extract, load, transform) method. The data can be stored on premise or off premise depending on the circumstances whether it is to be stored as historian data for later use or real time production data for immediate analysis. These are designed to hold data for a longer period of time before publishing or troubleshooting it further. It can be differentiated mainly into structured and unstructured.

1. Structured data: While comprising a relatively small sliver of the digital universe, it has been the main agenda of most of the organization today for better data management efforts. It is those data which are well organized in relational databases and can be queried by SQL. In an organization example of structured data may include financial records, department data, human resource, etc. which are beneficial to an organization internally. Also, real estate records, contact details, etc. which can be categorized under external structured data. The databases storing this type of structured data rely on some already predefined schema; that will ultimately sort data into tables comprising of rows and columns. Allowing SQL commands to be run across different databases for a specific data point, join data from varied table for different use cases comparison. It offers an ease of use that can be easily identifiable by id name, date, or time for analyzing.

2. Unstructured data: It comprises of the larger portion of digital universe. These
types of data are completely opposite to structured data and does not fit under the category of structured format and databases at all. They cannot be easily be adapted to relational databases. These types of data include free-form text, videos, captures, different sensor data, audio files. Unstructured data are often held in NoSQL relational databases, with an approach of “data first, schema later”. This attribute results in databases to store extensive range of data across organization which lead to the introduction of enterprise data lake in the market because when integrated with cloud the range of scalability multiplicated. If put in basic terms these data do not speak the same language but when extracting value from it requires work and is performed by big data analytics to the vast databases.

3.2.3 Data Catalog Layer

This is the third layer as we move up the architecture hierarchy where data can be inventoried; if tagged properly. This is an important layer as it stores all the meta data about the layer below it, and as the data are processed further. It can be tagged parallelly across all the required horizontal and vertical operations in an organization. This layer gives a bidirectional view of the data pipeline. It is in the position where it could be separated from higher hierarchies where new data inventories could be made according to the criticality of what you are doing and its importance to the business environment. A user can move down the hierarchy to get insight about the data sources for generating a new use case scenario or for different purpose. Similarly, a user can move up the hierarchy to get insight about the data sources which are used for reporting or for analyzing any inconsistencies. As different data sets are used for different purposes in and across organization, sometimes it becomes exhausting to find the correct data and its information about its handler in critical conditions.

3.2.4 Data Handling Layer

This is the fourth layer as we move up the architecture hierarchy. It involves process of collecting, cleaning and then storing, archived or disposed off in a secure manner; not only during the process of transformation but also after the conclusion about the information is generated and valued across different disciplinaries for insight. Basically, consolidating data into a single source from where it can be easily evaluated. This includes proper gov-
ernance policies satisfying the handlers of the data as well some recommendations for ease of use regarding glossary terms, ownerships, security. Data lake and data warehouse is an ideal resolution for this phase which compliments other integration tools and visualization tools for precise analysis and makes the data available to be viewed in many dimensions for different business scenarios keeping all the vulnerabilities check intact for precautions.

3.2.5 Data Analytics Visualization Layer

This is the fifth layer as we move up the architecture hierarchy, where data is readily available to users for different purpose, such as for analysis. Online analytical processing (OLAP) is an approach for multidimensional analytical process which allows the consumer of the data to investigate in many dimensions and decrease it precisely to two or three dimensions called cubes which is very specific or data mart which is little generic focusing on a single subject or functional organizational area and relatively few sources linked to one of business according to the demand of the user. The consumer of these data is business oriented to sales, finance, production, etc. It would be catastrophic for an organization if a data is misplaced when it is required the most during comparison process as its holds typ- ically the summarized data. Another purpose can be to provide graphical representation of information and data by visual elements representation like charts, graphs, maps, etc. The visualization tools proffer approachable trends and patterns in data. As the age of big data kicks into higher gear, visualization is also steering to make sense of the massive amount of data produced every day, curating into comprehensible visuals by highlighting the useful information. It is not an easy venture for combining data and analysis visuals together as a lot of effort goes in itself for cleaning, selection, delicate balancing between form and function and many more processes to compliment better decision making for the profitability of an organization.

3.3 Scenarios

As previously in the proposed system architecture, we saw that the data catalog layer provides the inventory about the data as its processes along the hierarchy. The main initial key considerations are the format of the data structure. It can be structured, unstructured and semi structured data, for example corporate relational data, CSV, JSON, logs, Images, etc. falls under these categories. The most common scenarios that evolved
in an organization due to these categories are.

1. Relational Database option in Azure.

2. Non-Relational Database option in Azure.

3. Composite Database option in Azure.

3.3.1 Relational Database Option in Azure

It refers to when an organization deals with structured data source format for reporting or analysis use case scenario, as shown in figure 25. Depending on the different purposes for which the data sources are used, Azure analysis service can be optional, and data can be directly published for reporting. In the SQL server family, the Infrastructure as a Service (IaaS) provisioning is SQL server virtual machine, whereas Platform as a Service (PaaS) provisioning are Azure SQL Database – Standard and Managed Instance (in preview), Azure SQL Data Warehouse[5].

![Organization dealing with structured data.](image)

Figure 25: Organization dealing with structured data.
In a Multi Parallel Processing (MPP) environment within an organization where scaling up or down, or pause, based on demand and integration with multi structured is required, the previous use case scenario is not feasible. Rather integration of Azure SQL database, Azure data lake store with Azure SQL data warehouse is feasible, as shown in figure 26.

![Organization dealing with structured data in a multi parallel environment.](image)

**Figure 26:** Organization dealing with structured data in a multi parallel environment.

### 3.3.2 Non-Relational Database Option in Azure

It refers to when an organization deals with data source format other than structured data, as shown in figure 27. While dealing with massive amount of data, it requires to be processed in a batch for different purposes or to be queried by users accountable. Integrated workspace with collaborative data science project ideats this scenario. The Infrastructure as a Service (IaaS) provisioning in this scenario are HDInsight, cloudera, MAPR, while Platform as a Service (PaaS) provisioning are Hive, Apache Spark and Azure Databricks.
3.3.3 Composite Database Option in Azure

It is an ideal scenario for any data driven organization dealing with different data source format in a multi platform environment. It is a combination of above two scenarios. Varied use case can be build for reporting, analysis purposes regarding the demand and necessity of an organization [95]. The initial key consideration for the catalog approach lead to an overview of different scenario possibilities in an organization dealing with varied data set aspect. Other considerations such as types of data ingestion pipelines, data size volumes, data movement and storage, architectural simplicity, information delivery and expectations of the users such as casual users, data analysts, data scientists, BI specialist has to be looked upon later in a process[5].
By the above scenario mentioned we can see that both vertical and horizontal components should compliment each other for information deliverables. Complexity can be reduced if information about the processes in the supply chain are shared. Certain regulations must be followed for the ease of sharability in a central platform, which is explained in the next section.

3.4 Data Governance

There is a great deal of talk in the recent years about how organizations are ideating over data – driven strategy. As human resources, operating budget, corporate facilities, etc. of an organization set certain protocols for managing different task. Similarly, data which is extensively present in an organization and precisely having huge range of audiences, does not have any rules on how we manage it. There was no support for data management until now, which gave rise to the evolution of data governance as a mainstream function and it has been growing strongly ever since[3].
3.4.1 Evolution

Predominantly it is about setting the rules of the road on how the data is to be managed across an enterprise and to unlock the value from that data. Initially, data governance came as data anarchy with nobody or everybody trying to manage the data. It was managed for local needs but with increase in data sharing it created impacts. When something happened in one place, it had an unexpected effect in some other places which basically stopped people from unlocking value of the data. So, the lack of management lead to data messes in enterprise and data governance was brought into picture to fix this situation as shown in figure 30, and is maturing paralleling to the enabling technologies.

Since then new companies were beginning to be formed that had data at the centre of their business models. Recent example like the acquisition of Whole Foods by Amazon; traditional supermarkets are also taking challenges in the business models. Since 2005, number of changing circumstances have given rise to new use cases and new protocols. We are at a new stage, where we got another generation of data governance.
3.4.2 Comparison of its Evolution Stages

In the above section we encounter the evolution of data governance which lead to modification of its policies from 1.0: data governance council to 2.0: data governance office to 3.0: agile governance, as shown in figure 31[57].

1. Data Governance Council

It was setup early in the early stage of 2005 when there was limited data set and little knowledge about technologies and its value. It was optional for different executives to invest their time on a council who directed to the formation of working groups of different projects accountable to address their own data needs, which lead to limited success as there was no resources specialized in data governance. In the long run it failed to generate value out of the data when new use cases began to progress.

2. Data Governance Office

Since the realization of the importance of data came to limelight, there was an urgent need of new set of rules and policies; which marshalled the introduction of centralized management system of top down approach focused on enterprise risk mitigation. An organization unit of dedicated staff specialists in data governance was set up to conduct policy development, process to manage the data in the enterprise. But it ceased to capture the
true needs of the audiences who actually work on the data.

3. Agile Governance

Due to the lack of focus on the audiences of the data who are actually working on the data steered the need of republic management system with bottom up approach to provide support to the staff who work with data, empowering them to contribute to the corpus of knowledge. Naturally, making a paradigm shift from tribal knowledge to shareable knowledge across a platform, where individual understand what they are working with.

Agile data governance provides that support to empower individual staff members to get the value of the data, but with empowerment comes responsibility. Expectations from staff would be to contribute to knowledge about the data and follow guidelines to protect data and make sure that the data which is going to be used in approved ways with the technology support, so that anybody can understand who is contributing and make sure people are contributing and crosscheck whether the guidelines being followed. So, if we think about it we could have a synthesis of the best practices for agile data governance. Which means that agile data governance takes the best from what we have already seen developed in the previous generations of data governance and puts it in a framework where its more supportive, after all it is known that data privacy and permitted use of data are still going to be important and different strategies in which that would be done can be accepted.
Agile data governance has certain principles such as pushing the policies and processes as close to the end-user as possible, as close to the point of use as possible that is just in time as it were and to enable that collaboration between users of the data. So, the central data governance office is still there, it is still producing policies and processes but rather than having those put in to some kind of SharePoint site where nobody knows how to get to them easily or to utilize them. But are available in a context where it is very close to when our data users are actually using the data. The data governance council in a form can there be as well the data governance office has to have executive support, it has to understand what business strategy it is. So, bottom up approach is better, but it does not need technology to support it because of its enterprise scope and the kind of range of capabilities that is asking people to be involved in like curating knowledge about the data, crowd sourcing for instance will be important in this kind of environment.

Figure 32: Differentiation between centralized and agile data governance.
It is true that there is emphasis on knowledge, but it cannot be expected all knowledge to be contributed by people. And after all we have vast ways of metadata out there about the data, just the system catalogs of databases, all kinds of metadata exist out there. So, as technology is evolving such as machine learning and automation, it distills the knowledge and adds it to what is being crowd sourced. So, data catalog is the principle platform for data governance support here and compliance becomes easier; for instance, data catalog can track who is using data. Here agile data governance actually goes a step before beyond central data governance. In this particular example of the policy it can track who is doing what with data and you get an idea of whether in fact the policy is being respected and that would not have been possible with centralized data governance.

3.5 Discussion

At initial stage of the thesis work at Boliden; the importance of cataloging approach in an organization was proposed to the meeting participant as mentioned in table 3.

<table>
<thead>
<tr>
<th>Meeting participant</th>
<th>Position</th>
<th>Work Field</th>
</tr>
</thead>
<tbody>
<tr>
<td>Andersson, Christer</td>
<td>IT Manager Data and Analytics Boliden</td>
<td>Target user “report builder and to support “data analyst”</td>
</tr>
<tr>
<td>Stenlund, Andreas</td>
<td>IT Mines Managers</td>
<td>management</td>
</tr>
<tr>
<td>Lloyd, Maria</td>
<td>IT Smelters Manager</td>
<td>management</td>
</tr>
<tr>
<td>van Ouwerkerk, Ivo</td>
<td>Architect and Service managers data and Analytics</td>
<td>Target user “report builder and to support “data analyst”</td>
</tr>
<tr>
<td>Lövgren, Daniel</td>
<td>Service managers data and Analytics</td>
<td>Target user “report builder and to support “data analyst”</td>
</tr>
<tr>
<td>Simon Emanuelsen ATEA (external)</td>
<td>report developer and project manager</td>
<td>Target user “report builder and to support “data analyst”</td>
</tr>
<tr>
<td>Sahlman, AlfBI NORDIC (external)</td>
<td>report developer and project manager</td>
<td>Target user “report builder and to support “data analyst”</td>
</tr>
<tr>
<td>Castehav, Stefan</td>
<td>Business responsible mines data &amp; analytics</td>
<td>Target user “data analyst” and management</td>
</tr>
<tr>
<td>Degerfeldt, David</td>
<td>Business responsible advanced analytics</td>
<td>Target user “data analyst” and management</td>
</tr>
</tbody>
</table>

Table 3: Participants of the meeting.

Value of the proposed architecture for Boliden was explained in brief referring to shareable system; where different users especially data analyst and report builders were targeted. Boliden confirmed the value of a data catalog in general, and agrees to proposed
scope and architecture. The proposal matches the pain points of Boliden, such as time used to find the data from the existing data assets; to be able to build reports or processes at later stages, which is explained in the implementation chapter.

3.6 Summary

In this chapter, we discuss about our proposed CATALYST architecture, which corresponds to the various processes that occur in an organization from data collection to data extraction to its visualization for getting value out of it. It follows by distinguishing them in five layers, where each layer has its own distinguishable attributes. However, consideration is given to the intermediate layer with agile data governance approach that fall within the scope of this study and its supporting factors. In the next chapter we describe the approach by implementing it in real life data assets scenario.
4 Implementation, Results and Evaluation

This chapter describes an implementation of the proposed catalog feature in an enterprise. The thesis delimitations are complaint to Boliden data landscape scenario only. We establish the experiment in two phase implementations. In former phase one implementation, we set up the proposed architecture environment by publishing real time production data sets from mining site to the platform. In the next phase of the experiment, we evaluate the proposal of cataloging layer approach. It will justify whether the proposed system is valuable for an organization or not, by finding the desired data with less complexity. With the hypothesis build around our proposal; recommendations are suggested to build a shareable platform for Boliden.

4.1 Prototype Implementation

As discussed in cataloging needs of chapter 2, due to complexity in the approach of vertical and horizontal integration of information; insight to a data is lost easily. We will gauge our approach in this phase one implementation. For our implementation, we deal with real time production and reports data assets (mostly outsourced to Atea company) from Aitik, Renstrom, Kankberg, Kristinebrg and Garpenberg mining sites. Figure 33,34,35 are the few examples of data asset which we will be dealing with for the demonstration of our approach. Figure 33 denotes information about the timestamp of a machine’s activity from Kristineberg site. Figure 34 denotes the same information collected about the timestamp of a machine’s activity from Garpenberg site. Figure 35 denotes information about the blast activities and metals generated from Garpenberg site in one day.

Figure 33: GSACTIVITYDATAKINDVALUES: Excel file machine activity data details from Kristineberg.
It can be assumed by seeing the figures that large amount of data is produced from different locations. The challenge for mining, automation and process optimization are key driver for productivity increase. Data driven insights and decision have an unused potential when it comes to enabling production planning and monitoring; with higher automation, increase efficiencies based on data, identification of bottlenecks to determine process optimization or automation potential. Safety is always a priority and here data and analytics can enable to move away people from dangerous places or situations. The assumption is that by building up a Boliden data and analytics IT capabilities platform, will enable to scale and operate implemented use case more efficiently but also be faster with helping our engineers and researchers on the value driven data insight and decision. For this to happen, initial context must be given to the data as it is processed along the different layers of the proposed architecture.
4.1.1 Data Consumers

In the traditional system, discovering existing data sources has been purely based on tribal knowledge. For Boliden, that want to get the most from their information assets, this approach presents numerous challenges. It can be noticed from the data sets, as shown in previously in the section. The user might not know that a data source exists if generated at different locations, as there is no central sharable location where the data sources are registered. For any enquiry about an information asset, a user must locate the expert accountable for that data; which may end in chaos if there is no proper input. There is no credibility when reusability of data sources scenario arises as the perspective of the usage might be different for varied user demand.

4.1.2 Data Producers

In addition to the data users that face the previously mentioned challenges, users who are responsible for producing and maintaining information assets face challenges in the similar fashion. Descriptive metadata is often ignored as it can be a tedious job. Documentation without sync with data sources at each process maybe not trustworthy to consume. Restricting access to data sources is a taxing challenge.

When data consumers’ and data producers’ challenges are combined, can impose a significant hurdle for companies who want to venture into self-service platform where understanding of enterprise data is important. Here the data catalog service plays a pivot role. Once the data sources are published in the cloud-based service, the metadata could be enriched parallelly along the process. The metadata could be added by any user in an organization which makes it approachable by any consumers across the organization too.

4.1.3 Metadata Enrichment

Distinct users have different stance regarding the same data sources. The system administrator and the owner have knowledge about the Service Level Agreement (SLA) for the servers and the services that host the data source or accessing the request access for users, the analyst has idea about the data processes that take place in the context of the business operations, the data steward have better understanding about the attributes of the data sources and its mapping with the business models. Proper organization of metadata is
lacking which could be established, if proper annotation or tagging are available for each perspective of the consumer and the producer. The different annotations endorsed by the data catalog are mentioned below

<table>
<thead>
<tr>
<th>Annotation</th>
<th>Notation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Friendly name</td>
<td>Friendly names could be provided when the underlying object name and the definition of the data asset is complex.</td>
</tr>
<tr>
<td>Documentation</td>
<td>Documents like attachments, images could be provided to make better observation about the data asset</td>
</tr>
<tr>
<td>Description</td>
<td>Description could be provided by the user regarding its own perspective about the data asset</td>
</tr>
<tr>
<td>Experts</td>
<td>Expert name could be provided initially when the data assets are published, so that he can create a user group in his domain or contact point of reference when answers for particular questions could not be provided by annotations. The expert can take the ownership and give access to user group when the situation demands for it.</td>
</tr>
<tr>
<td>Tags (user tags)</td>
<td>User defined tags could be provided to categorize the data assets better as data assets are processed along the data pipeline for an assessment</td>
</tr>
<tr>
<td>Tags (glossary tags)</td>
<td>Centrally defined tags could be provided, so that better categorization could be provided and everyone in an organization uses same business taxonomy.</td>
</tr>
<tr>
<td>Request access</td>
<td>Request information details could be provided as to who can access the information about the data access for security reasons.</td>
</tr>
</tbody>
</table>

Table 4: Annotations and its definition.

By using these annotation, the existing enterprise data assets could be better organized. This would give a consumer better understanding and to find a data asset in this huge organizational big data scenario.

4.2 Testbed

For our experiment, initial set up was done by creating euwbdpgrg0001 main resource group in Microsoft Azure platform, followed by generating sub resource group bolentsrvdb001 with type SQL server and bolentsrvdb001 with SQL database for implementing data catalog service, as we can see it in the table 5, 6, 7 and 8 for the reference.
<table>
<thead>
<tr>
<th>Type</th>
<th>Resource group</th>
<th>Subscription</th>
<th>Subscription ID</th>
<th>Tags</th>
</tr>
</thead>
<tbody>
<tr>
<td>Resource group</td>
<td></td>
<td></td>
<td>a257378b-0c31-44e0-94c0352ec49d51a3</td>
<td>projectnumber costcenter activitynumber technicalcontact partner purpose expectedlifetime createdby createdate</td>
</tr>
<tr>
<td>eubwdbpgr0001</td>
<td>Boliden-Production</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5: Type - Resource group

<table>
<thead>
<tr>
<th>Type</th>
<th>Resource group</th>
<th>Location</th>
<th>Subscription name</th>
<th>Subscription ID</th>
<th>Server name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL server</td>
<td>eubwdbpgr0001</td>
<td>North Europe</td>
<td>Boliden-Production</td>
<td>a257378b-0c31-44e0-94c0352ec49d51a3</td>
<td>eubwdbpgr0001.database.windows.net</td>
</tr>
<tr>
<td>bolentsrvdb001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6: Type - SQL server

<table>
<thead>
<tr>
<th>Type</th>
<th>Resource group</th>
<th>Location</th>
<th>Subscription name</th>
<th>Subscription ID</th>
<th>Server name</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQL database</td>
<td>eubwdbpgr0001</td>
<td>North Europe</td>
<td>Boliden-Production</td>
<td>a257378b-0c31-44e0-94c0352ec49d51a3</td>
<td>eubwdbpgr0001.database.windows.net</td>
</tr>
<tr>
<td>bolentsrvdb001</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7: Type - SQL database

<table>
<thead>
<tr>
<th>Type</th>
<th>Resource group</th>
<th>Location</th>
<th>Subscription name</th>
<th>Subscription ID</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data Catalog service</td>
<td>eubwdbpgr0001</td>
<td>North Europe</td>
<td>Boliden-Production</td>
<td>a257378b-0c31-44e0-94c0352ec49d51a3</td>
<td>Azure Data Catalog Portal <a href="https://www.azuredatacatalog.com/">https://www.azuredatacatalog.com/</a></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 8: Type - Data Catalog service
It is verifiable that the data produced from the mining site are available in a decentralized manner across the organization, and stored either in the on-premise or off-premise databases. According to the demand, these data were used to generate reports or for analysis purposes. If we refer to figure 7; it denotes that the data sets were used from the databases directly. This prevented the supply chain of Boliden from further enhancement, as data could not be reused when required for further analysis. Data isolation became more prevalent and required more resources to build a report or analysis from scratch; sometimes leading to duplicated data which is not cost efficient. To prevent all these challenges for our implementation we published data related to production of the mining site initially from the on-premise databases at different geo-location to azure cloud database; refer to figure 36.

Figure 36: Data catalog database dashboard.
4.2.1 Inventory of Data Assets

For the next step bolentsrvdb001 was used as the data source for data catalog platform, as shown in figure 37. After successful registration to the source and integrating with client application tools (Excel, Excel (Top 1000), Power Query), data sources were published to the catalog. The published data in the catalog service gave a centralized view of all the data from different geo-location of mining sites. Not only real time data from the production site but also the data that were directly used for final reporting purpose, from different geo-location were provided to this centralized platform. In the below figure 38 and 39, we can perceive that data from device layer, data source layer; referred as gannt data and visualization layer of the proposed architecture are shared in a central platform.
Figure 38: Published gannt data to data catalog, page 1.

Figure 39: Published gannt data to data catalog, page 7.

Figure 40 and 41, verifies that final budget report from Garpenberg were also published to catalog for our implementation; from the client application Microsoft Power BI tool.
Power Query is the Microsoft Data Connectivity and Data Preparation technology. It is natively integrated in Microsoft products such as Microsoft Power BI, Microsoft Excel just to name a few. It enables report builders to access data; stored in decentralized manner; to shape a report as per the requirement. Figure 42, shows the complexity of generating report at visualization layer itself. Initially, the required data were collected from various sources such as the excel, osisoft, centuri, datablast and ipak. From the collected data, only the required information for generation of a specific report were extracted, by using different queries. These budget reports which were built were published to catalog for our implementation.
By this implementation; we tried to publish the data sources from the existing data assets to a centralized platform and tagging them for better classification. Implying to the existing challenges for the vertical and horizontal integration in an organization, it showcases that each layer’s complexity is hidden, by bringing simplicity to the procedure of information sharing. But as volume of data sources increases in a central platform, it is important that this approach should be efficient enough for creating output with less duplicity, time and resources. This we explore in the next section.

4.3 Evaluation

For this experiment we set up an environment with two use case scenarios of catalog perspective and non-catalog perspective, as shown in the skeleton architecture figure 43.
In the catalog perspective, the heterogenous data assets with datapoints were categorized properly by tags such as GANNT, GANNTREPORT and BUDGETREPORT. Sub tags with type raw data, kristinebergreport visualization, renstromreport visualization and aggregatedmetal1report visualization and aggregatedmetal2report visualization. Whereas in the non-catalog perspective, the heterogenous data assets with datapoints were not categorized by tags at all, as shown in the figure 44.

We delimitated our datapoints to maximum 1000 rows, as maximum of the data assets used for this evaluation had maximum of 30,000 datapoints. Table 9 refers to the delimitation consideration for implementation.
4.3.1 Catalog Perspective

The production data that will be used in the analysis process and well as for generating budget report were established centrally with tags; where the data consumers and the data producers could easily find the data asset. In this implementation, we demonstrated the efficiency of our proposed approach with SQL query hits and recorded the individual time required for each subquery for a complete process to be done. Figure 45 is the snippet code for our evaluation.
The request was initiated for the GANNT details, as shown in the figure 46. It generated queries and subqueries to get all the existing production data assets with the tags from the whole data assets, as shown in figure 47. We record all the individual query hits and the time take for each query to process for the evaluation of our system. Similarly, two more requests were initiated for generating reports, and the time taken for each query hits and data counts were recorded; as shown in tables 10, 11 and 12.
Figure 47: Query and subqueries generated for the request made.

<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>1.801875</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>0.958993</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 3</td>
<td>1.964078</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 4</td>
<td>1.274007</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 5</td>
<td>1.326194</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 6</td>
<td>1.205246</td>
<td>383</td>
</tr>
<tr>
<td>Subquery 7</td>
<td>1.452431</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 8</td>
<td>1.38085</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 9</td>
<td>0.386115</td>
<td>11</td>
</tr>
<tr>
<td>Subquery 10</td>
<td>1.059055</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 11</td>
<td>1.472178</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 12</td>
<td>1.430588</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 13</td>
<td>0.395284</td>
<td>60</td>
</tr>
<tr>
<td>13</td>
<td>16.716885</td>
<td>10454</td>
</tr>
</tbody>
</table>

Table 10: Individual query hit details for gannt data.

<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>0.682667</td>
<td>199</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>1.1721</td>
<td>409</td>
</tr>
<tr>
<td>Subquery 3</td>
<td>0.894237</td>
<td>409</td>
</tr>
<tr>
<td>3</td>
<td>2.749004</td>
<td>1017</td>
</tr>
</tbody>
</table>

Table 11: Individual query hit details for gannt report.
4.3.2 Non-Catalog Perspective

The production data that will be used in the analysis process and well as for generating budget report were not established centrally with tags. This scenario was the contrast to the previous catalog prospective. The production data as well as the reports were located at different geo-location of the mining site. In this implementation, we demonstrate the query hits; refer to figure 48, 49, 50 and record the individual time required for each query hits and data counts for each subquery; refer to tables 13, 14 and 15.

<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>1.518018</td>
<td>610</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>1.168574</td>
<td>610</td>
</tr>
<tr>
<td>2</td>
<td>2.686592</td>
<td>1220</td>
</tr>
</tbody>
</table>

Table 12: Individual query hit details for budget report.

Figure 48: Code snippet for non-catalog evaluation.
Figure 49: Request send for data in non-catalog scenario.

Figure 50: Query and subqueries generated for non-catalog scenario.
<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>2.296597</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>1.24756</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 3</td>
<td>1.17739</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 4</td>
<td>1.783891</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 5</td>
<td>1.528596</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 6</td>
<td>0.729918</td>
<td>383</td>
</tr>
<tr>
<td>Subquery 7</td>
<td>1.749335</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 8</td>
<td>1.391075</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 9</td>
<td>0.634006</td>
<td>11</td>
</tr>
<tr>
<td>Subquery 10</td>
<td>1.683395</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 11</td>
<td>2.560002</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 12</td>
<td>3.000069</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 13</td>
<td>0.686193</td>
<td>60</td>
</tr>
<tr>
<td>Subquery 14</td>
<td>1.063846</td>
<td>199</td>
</tr>
<tr>
<td>Subquery 15</td>
<td>1.26766</td>
<td>409</td>
</tr>
<tr>
<td>Subquery 16</td>
<td>0.772937</td>
<td>409</td>
</tr>
<tr>
<td>Subquery 17</td>
<td>2.336091</td>
<td>610</td>
</tr>
<tr>
<td>Subquery 18</td>
<td>1.522249</td>
<td>610</td>
</tr>
<tr>
<td>18</td>
<td>27.43081</td>
<td>12691</td>
</tr>
</tbody>
</table>

Table 13: Individual query hit details for gannt data.

<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits in Non-Catalog (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>2.873802</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>1.936284</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 3</td>
<td>1.890765</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 4</td>
<td>1.365104</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 5</td>
<td>1.271681</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 6</td>
<td>1.546549</td>
<td>383</td>
</tr>
<tr>
<td>Subquery 7</td>
<td>1.522559</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 8</td>
<td>1.401299</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 9</td>
<td>1.128229</td>
<td>11</td>
</tr>
<tr>
<td>Subquery 10</td>
<td>0.691523</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 11</td>
<td>3.335177</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 12</td>
<td>2.277386</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 13</td>
<td>1.547573</td>
<td>60</td>
</tr>
<tr>
<td>Subquery 14</td>
<td>0.804319</td>
<td>199</td>
</tr>
<tr>
<td>Subquery 15</td>
<td>1.057497</td>
<td>409</td>
</tr>
<tr>
<td>Subquery 16</td>
<td>0.810666</td>
<td>409</td>
</tr>
<tr>
<td>Subquery 17</td>
<td>1.325466</td>
<td>610</td>
</tr>
<tr>
<td>Subquery 18</td>
<td>0.685124</td>
<td>610</td>
</tr>
<tr>
<td>18</td>
<td>27.875103</td>
<td>12691</td>
</tr>
</tbody>
</table>

Table 14: Individual query hit details for budget report.
4.3.3 Result and Discussion

The comparison was observed for all the three categories of the query made in both the catalog and the non-catalog perspective. GANNT data was the first category of the query request made. As shown in figure 51; it was observed that in catalog perspective, there were only 13 query hits whereas in non-catalog perspective, there were 18 query hits made. Aggregated time required to process all the query was more in non-catalog perspective too, with maximum approximation of 2 millisecond in catalog and 3 milliseconds for non-catalog.

<table>
<thead>
<tr>
<th>Query hits</th>
<th>Time required for individual query hits in Non-Catalog (milliseconds)</th>
<th>Individual data counts</th>
</tr>
</thead>
<tbody>
<tr>
<td>Subquery 1</td>
<td>2.673882</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 2</td>
<td>1.736284</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 3</td>
<td>1.603348</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 4</td>
<td>1.291635</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 5</td>
<td>1.230243</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 6</td>
<td>0.661509</td>
<td>383</td>
</tr>
<tr>
<td>Subquery 7</td>
<td>1.130489</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 8</td>
<td>0.431955</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 9</td>
<td>1.45278</td>
<td>11</td>
</tr>
<tr>
<td>Subquery 10</td>
<td>1.380142</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 11</td>
<td>0.514467</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 12</td>
<td>0.822302</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 13</td>
<td>0.821949</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 14</td>
<td>0.774699</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 15</td>
<td>0.923151</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 16</td>
<td>0.829534</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 17</td>
<td>0.829935</td>
<td>1000</td>
</tr>
<tr>
<td>Subquery 18</td>
<td>0.685134</td>
<td>1000</td>
</tr>
<tr>
<td><strong>18</strong></td>
<td><strong>22.794487</strong></td>
<td><strong>12691</strong></td>
</tr>
</tbody>
</table>

Table 15: Individual query hit details for gannt report.

Second category of the query request was made for gannt reports, as shown in figure 52. It was observed that in catalog perspective, there were 3 query hits with maximum of 1.2
milisecond captured for the individual request. Whereas in the non-catalog perspective, with maximum of 3 millisecond captured for the individual request.

![Figure 52: Comparison of the same query made for gantt report.](image)

Third category of query request was made for Budget report, as shown in figure 53. It was observed that in catalog perspective, 2 query hits with maximum proximity of 1.5 milliseconds captured for individual query hits. Whereas in non-catalog scenario, 18 query hits with maximum proximity of 4 milliseconds captured for time required to be processed.

![Figure 53: Comparison of the same query made for budget report.](image)

After the outcome of both the prospective; comparison was made between them. From the figure 54, it could be evaluated that the graph with non-catalog perspective implementation (marked in blue color) not only had a greater number of query hits but also took more time for individual sub queries to be processed. While in the catalog perspective (marked in yellow, green and red) all the three query requests were processed with a smaller number of query hits and less time period compared with the former.
When comparison of the aggregated query time were made between the approaches; it verified that catalog implementation still showed better result, as shown in the below table 16 and the figure 55.

<table>
<thead>
<tr>
<th>Data assets</th>
<th>Non-Catalogue</th>
<th>Catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budget Report</td>
<td>27.875103</td>
<td>2.686592</td>
</tr>
<tr>
<td>Gantt Report</td>
<td>22.794487</td>
<td>2.749004</td>
</tr>
<tr>
<td>Gantt</td>
<td>27.43081</td>
<td>16.718885</td>
</tr>
</tbody>
</table>

Table 16: Aggregated query time.

Figure 55: Comparison of query time between catalog and non-catalog approach.
Aggregated data count was also compared between the two approaches for further evaluation, and both the approaches showed huge difference. It proved that the catalog approach implementation was promising as shown in the below table 17 and the figure 56.

<table>
<thead>
<tr>
<th>Data assets</th>
<th>Non-Catalog</th>
<th>Catalogue</th>
</tr>
</thead>
<tbody>
<tr>
<td>Budget Report</td>
<td>12691</td>
<td>1220</td>
</tr>
<tr>
<td>Gantt Report</td>
<td>12691</td>
<td>1017</td>
</tr>
<tr>
<td>Gantt</td>
<td>12691</td>
<td>10454</td>
</tr>
</tbody>
</table>

Table 17: Aggregated data counts.

![Comparison of Data Counts between Catalog and Non-Catalog approach](image)

Figure 56: Comparison of data counts between catalog and non-catalog approach.

From the implementation in phase one, where we set up the catalog platform approach; to the importance of tagging each data source from the existing data assets in phase two of the implementation, where achieved in this section. Figure 57, finally verifies by showing the contrasting difference in both the approaches. In the catalog perspective, the average query time as well as the average data count was reduced to less than 50 percent. The average query time was reduced from 59.813 milliseconds to 11.009 milliseconds in the catalog perspective. Also, the average data count was reduced from 12,691 to 5721.7 in the catalog prospective. This proved that our proposed CATALYST architecture was more efficient in terms of the data isolation challenges faced by the organization, to enhance the capabilities of the supply chain.
4.4 Limitation

While setting up the actual platform with its evaluation, it was noticed that the tag insertion and publishing of the data assets were done manually. When this approach will be mimicked to a large-scale base, where data will be integrated to a central platform with data assets flowing from all the different layer as proposed in our architecture; can end up in complications. There can be a fault in tagging and publishing data sources centrally by humans. When this situation occurs, the misplaced information about the data assets can cause vulnerabilities in the whole process. This will demand for a creation of new data sets from scratch, and end up with duplicate data in the system, wastage of resources and longer time period to get insight from data. So, resources should be hired to do the task in compliance with the expert’s advice.

4.5 Recommendations

It was noticed that our proposed CATALYST architecture was more efficient if tags were provided to the data assets. The tags that were used in the figure 58 are as follows:
1. Data Layer - This tag suggests that the data information showcased in the data catalog are raw data from the actual device layer. This will help to gather segregated data from different mining sites.

2. Department - This tag suggests that the data information showcased in the data catalog are for a particular department accountable in an organization such as for finance department, marketing department, IT department, etc. This will help the users of the data to target on a particular data assets rather than getting lost by it.

3. Locations - This tag suggests that the data information showcased in the data catalog are from a particular mining sites. This will help the target users to fetch the data directly that are required from a particular site rather than contacting different individuals regarding that data.

4. Reports - This tag suggests that the raw data were processed upon and the information flow is from data visualization layer to the data catalog. This will help the department accountable users to make better decisions.

5. Outsource - This tag suggests that particular data were outsourced to and from from other organizations. This will help to keep track about a data assets when required for later process.

6. Prioritized - This tag suggests that a particular data published in a data catalog are of more importance. This will benefit to segregate data assets more efficiently.

7. Multitenancy - This tag suggests that a particular set of people are initially granted read access except the expert designated for a data asset in the data catalog. This will help to trace whether the data being used are with specific users.

8. Expected Multitenancy - This tag suggests that more number of people will be granted read, write or both access by the expert in the data catalog. This will help to trace whether the information about a data assets are not misused.
It will enhance Boliden system in the long run, to find a data asset from the existing enterprise data sources. Few other recommendations were made for later use as well, and Boliden confirmed its viability in the long run as suggested in figure 59.

As an organization has to deal with a lot of data. It is always better to get insight from the data within a short time period. This is only possible if the data assets are segregated with proper tags initially as well as at later stages of the value chain when required. Initially, the approach of a proper agile data governance, where the input of the data producers and consumers is really valuable, should be supported. Common terminologies has to be set up because a particular data can have different characteristics depending on its audience. Not only with the terminologies but also to get the most value from their information assets; it is important to determine which copy of data should be registered and published to a catalog depending on the needs and requirements.

Once all the initial criteria are resolved as mentioned previously, the next phase of tagging becomes critical. As mentioned in figure 24, the attribute of a data changes in each layer. This leads to divergent use case scenarios with possible tags as mentioned in figure 59.
Figure 58: Recommendations for tagging in the catalog for Boliden.
4.6 Summary

This chapter has provided a detailed description of our proposed system implementation by proper evaluation and comparison of its results. The verification on the efficiency of our approach was investigated and approved by Boliden. Limitations were found and few recommendations were made considering Boliden for leveraging their analytics and reporting capabilities. This will help the organization for organizing the existing enterprise data assets in the present as well as in the future projects by decreasing the gap of data isolation.
5 Conclusion and Future Work

This thesis provides a proposal to the challenges being faced in today’s data driven organization, where insight to the enterprise data assets is lost due to the huge varied organization data landscape. Initial analysis of the current situation of an organization dealing with existing data sets was done, which leads to our research motivation to solve this problem set.

Our work bridges the gap of tribal knowledge to shareable knowledge by bringing all the department accountable in and across organization to operate functionally. Exploration about the enabling technologies and the related work ongoing in the commercial applications and industry were done; to investigate more about the challenges and the needs to resolve it. This leads to the proposal of CATALYST: A Cloud-Based Data Catalog System for a Swedish Mining Company architecture later in our report, by considering the format of the data structure with different architecture scenarios in Microsoft Azure cloud platform. Our approach also looked upon, on how important is agile data governance management approach for our proposed architecture.

Finally, our work was justified by showing that our CATALYST architecture proposal is more efficient, when data assets are published in a systematic manner before and after processing it. The result showcased that data isolation could be prevented if our approach was implemented in real world scenario at Boliden. As our experiment was performed in a small scale scenario; it could be mimicked for a larger scale scenario as well, when more enterprise data assets are published to data catalog. This will lead to a system for further development and improvement. We looked into shareable, reusability and lessening of the data isolation factors. More factors could be looked into as the system matures. Recommendations were made accordingly for the future work, when more processes will be added up to the supply chain for better implementation.
References


Chris Patil and Vivian Siegel. “Shining a light on dark data”. In: Disease Models & Mechanisms 2.11-12 (2009), pp. 521–525. ISSN: 1754-8403. DOI: 10.1242/dmm.004630 eprint: [http://dmm.biologists.org/content/2/11-12/521.full.pdf](http://dmm.biologists.org/content/2/11-12/521.full.pdf) URL: [http://dmm.biologists.org/content/2/11-12/521](http://dmm.biologists.org/content/2/11-12/521).


[77] Andreas Tolk and James A Muguira. “The levels of conceptual interoperability model”. In:


