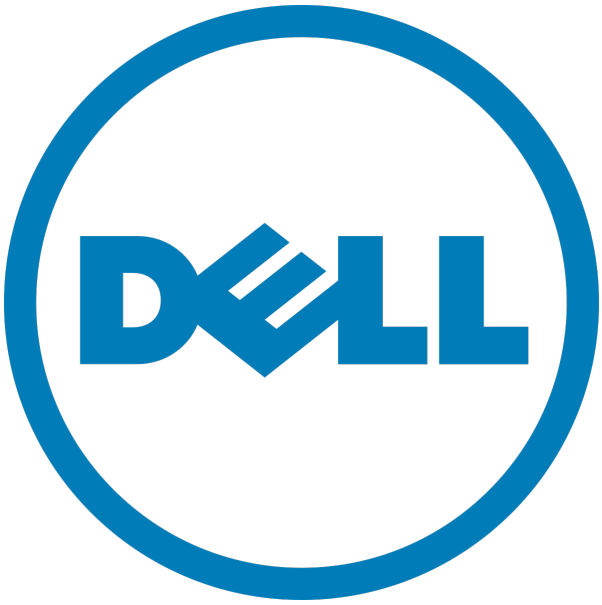


Building a 360-degree Customer View

By Jonas Hellevang

MBD S1 October 2019

Table of Contents

Why is This a Big Data Problem?

"Big data is when the data itself is a part of a problem[1]" - Presentation from 2013 Dell Enterprise Forum

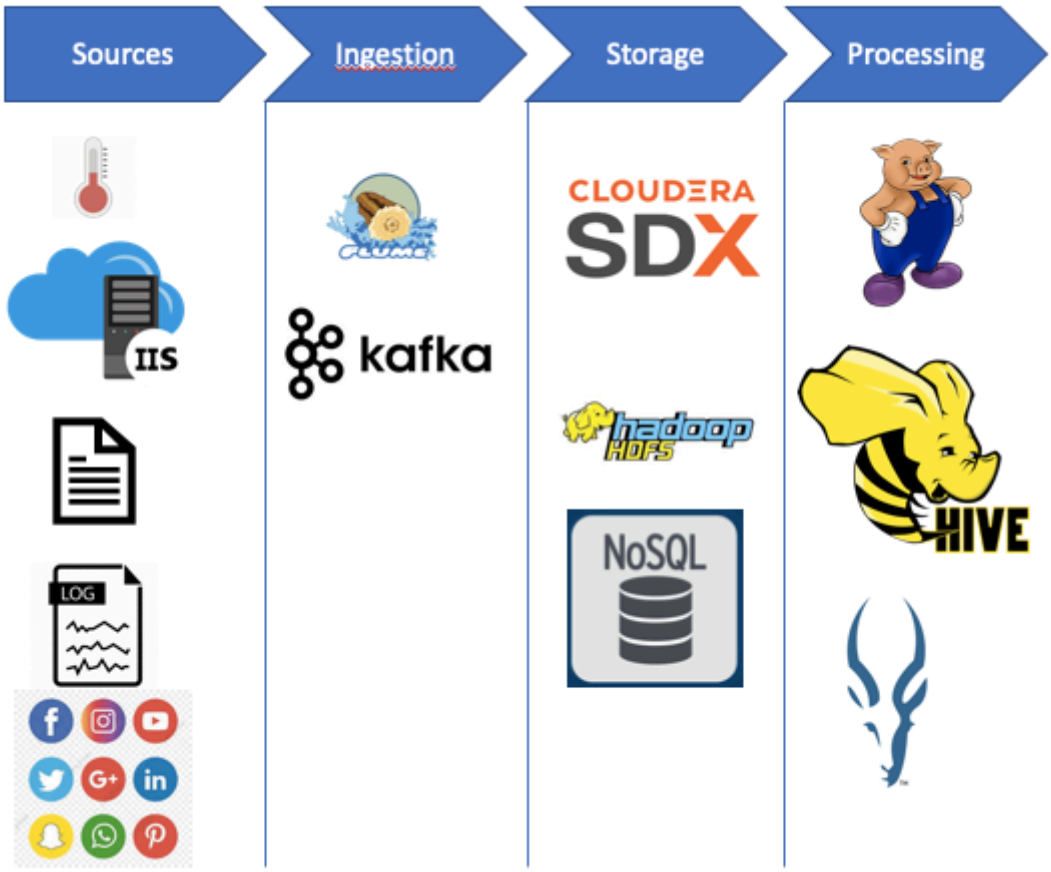
As the world grows larger and we collect more data, systems demand more power, speed and variety. Therefore, it is extremely important for companies to always deliver, improve and continuously follow the trends of technology. This require companies to have top of the shelves supply chain operational systems, giving the users of the system real-time insights into all aspects of the process[2]. Having real-time updates of all the information, at all times, is crucial to the logistics and delivery of the products and will surely increase the profit of the company.

Dell saw that they in only 5 years will increase the amount of data produced each year by 8-10 times, making the first V of big data, volume, a key problem. How will they be able to store all this data? When the volume is going up, it is also important to maintain or increase the second V, Velocity. How can they make sure customers with increasing demands can have real-time insights into their data? As Dell is tackling the new approach of prediction instead of explaining what happened, Dell made a decision to tackle this big data problem, setting these minimum criterions for what the solution should do*:*

1. *Store high-volume, multi-structured and high-velocity data sets over long periods of time*
2. *Process streaming data in near real-time, enabling on-the-spot issue detection and resolution*
3. *Converge data silos into one unified environment*[3]

[1] https://www.slideshare.net/lhrc_mikeyp/ai10-optimizing-poweredgehadoopvfnl/2
[2] https://www.cloudera.com/solutions/technology.html
[3] https://www.cloudera.com/about/customers/dell.html

Data Architecture



Data Ingestion

Data Sources

The sources of the data are many and different, which is one of the reasons why Dell needed this change. Dell had a lot of unused unstructured data, but with the new tools they can access all of it within seconds. The datafiles come in all shapes and sizes at dell, from audio tapes with clients converted into text, to sensors that measure temperature and machine data. Other data sources are social media, weblogs, PDF's, application data, data from repair centers, data from vendors and suppliers, and so many others. The data is mostly text, tables etc., giving you a huge variety of needs on how to deal with these data types.

Ingestion Tools

Data ingestion is the method of loading, importing, cleaning and preparing data to be loaded into the data storage. Data ingestion can be in real-time, all the time, once in a while etc. There are a lot to choose from, but Cloudera and Dell have chosen to work with Apache Flume and Apache Kafka, working with them simultaneously.

Apache Flume

Apache Flume is great for streaming huge amount of data, and that includes huge amounts of logs. It's efficient, reliable, and built for the Hadoop scale. It is easy to connect to, and works well with HBase, which is a NoSQL database that works on top of Hadoop.

Apache Kafka

Apache Kafka is an open source distributed streaming platform. Kafka is reliable, scalable, secure and fast. Kafka was developed by LinkedIn and is an excellent tool to use for real-time data pipeline. Dell and Cloudera are using it just for that, giving real-time feedback to the user. Dell and Cloudera have integrated Flume and Kafka together, making it a more powerful and better tool.

Data Processing

Data Storage

According to Cloudera, their Cloudera Operational Database delivers the next generation database. It is fast, it is easy, it is scalable, it is available, it is secure, and it works well with the other components.

Apache Hive

Apache Hive is a data warehouse software that manages big datasets. Data warehouses are structured, expensive for huge data volumes, and is designed to maximize scalability, extensibility, etc. Hive was built on top of Apache Hadoop by Facebook.

Apache HBase

HBase is a NoSQL database that connects to the Cloudera platforms. It is modeled after Google and runs on top of the HDFS. It is not a direct replacement for SQL, but it provides most of what is needed.

Apache Impala

Apache Impala lets you query data no matter where it is stored. It lets you use the SQL language – in real-time. Impala directly access the data from MapReduce, giving you a faster result than you normally would get.

Apache Pig

Apache Pig uses a language called Pig Latin and will be able to do its job in MapReduce or Apache Spark. Apache pig can be extended using other programming languages, and is as the other Apache tools here able to run on Hadoop.

Project results

The solution gives Dell a platform that works well with different number of workloads, sizes, datatypes, using the Dell/Cloudera cooperation as basis, and the Cloudera support for solving issues that appears. The solution drives insight into the aspects of:

- Marketing
- CRM
- Supply chain operations
- Hardware prototypes validations
- HR

Cloudera supplies what Hadoop and Dell provides, such as the great tools of Cloudera Navigator, Cloudera Manager and Cloudera Search.

What has this project led to? Most importantly the configurations have worked. All tools are working well together, and Dell can get the information they need, when they need it. It is also important to see the value that the customers get. Dell is able to provide any information the customer asks for and is able to help implement whatever is missing. Even though it is a huge step to create something yourself, even with such big companies as Cloudera and Dell, you should always consider using the open sources available in the ecosystems of Hadoop.

Sources:

https://www.predictiveanalyticstoday.com/data-ingestion-tools/

https://www.slideshare.net/lhrc_mikeyp/ai10-optimizing-poweredgehadoopvfnl/2

https://www.cloudera.com/solutions/technology.html

https://www.cloudera.com/about/customers/dell.html

https://mapr.com/products/apache-hive/

https://www.cloudera.com/content/dam/www/marketing/images/logos/cloudera/cloudera-newco-logo.png

https://upload.wikimedia.org/wikipedia/commons/thumb/4/48/Dell_Logo.svg/1200px-Dell_Logo.svg.png

https://www.em360tech.com/the-total-economic-impact-of-dell-cloudera-apache-hadoop-solution-accelerated-by-intel/