Implementing a Data Lake or Data Warehouse Architecture for Business Intelligence?

Business Intelligence (BI) is about **using data of yesterday and today to make better decisions about tomorrow**. It can be understood as **the function which ensures that raw data is transformed into meaningful information that provides insights and enables decision-making**. If we think about what companies have been trying to do for a long time, very often we hear about **investing in technology and tools that are supposed to solved business problems with data and analytics**. BI, however is less about technology and tool and more about using data , technology and tools to **create business insights**.

In a nutshell, **BI involves gathering functional business requirements and translating it into technical solutions by designing data models, doing the ETL — Extract, Transform and Load process to bring data from operational source systems to an analytics/destination database**, which can **be used to visualize information in the form of a real-time automatic dashboard**. This is done in a business to make informed decisions based on the past data rather than on a " Gut instinct".

What is the ultimate output of BI?

1. BI provides a single version of truth

The story of **A Single Version of Truth is the story of a data dictionary and data source**, which is **all about the meaning of data** which should be agreed

across an enterprise. And not less important, where to source the data. Let me give you an example. I used to work for this bank who can not answer a simple question such as: How many unique customers does it have? Depending on who you asked it, you will receive a different number. **Having an unique source of truth is therefore a must**. Instead of going through multiple spreadsheets with different numbers and doing reconciliations, BI product provides users a real-time automatically updated and consistent report.

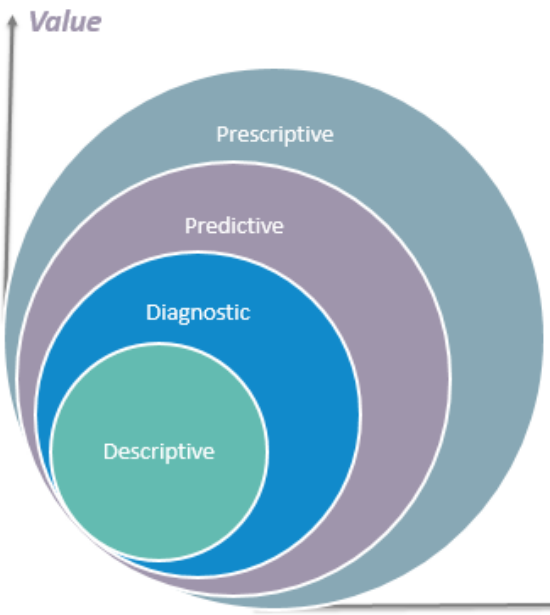
2. BI provides descriptive analytics

Business intelligence is descriptive because it **tells you what' s happening now and what happened in the past**. It informs **how a company is doing with regard to its set KPIs and metrics**. For instance, BI product can tell a manager **how the company is making sales and how far it is from reaching its set goals**. This information is usually provided in the form **of dashboards that include bar graphs, line charts** and the like, which gives users the most important information at one glance.

3. BI provides diagnostic analytics

Diagnostic analytics is about giving in- depth **insights and answering the question**: **Why something happened?** BI dashboards **provides drill- down functionality — which mean, from high- level overview, users can slice- and- dice the information from different angles, to the very details to figure out why things happened**.

Bi does not tell you what's going to happen in the future.



Descriptive: ***What's happening in my business?***

- Comprehensive, accurate and live data
- Effective visualisation

Diagnostic: ***Why is it happening?***

- Ability to drill down to the root-cause
- Ability to isolate all confounding information

Predictive: ***What's likely to happen?***

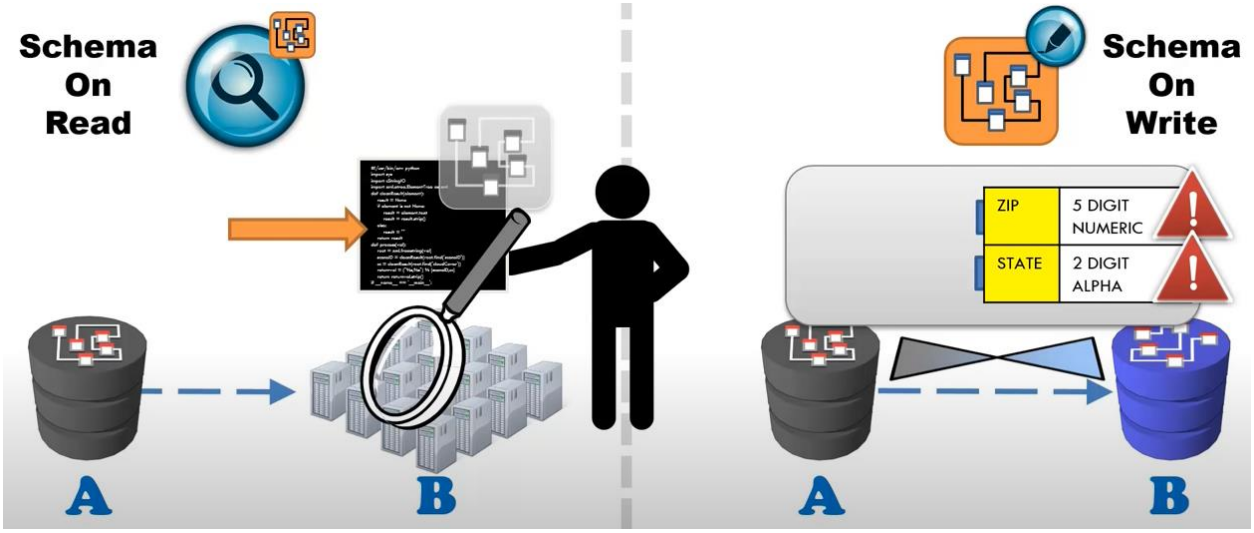
- Business strategies have remained fairly consistent over time
- Historical patterns being used to predict specific outcomes using algorithms
- Decisions are automated using algorithms and technology

Prescriptive: ***What do I need to do?***

- Recommended actions and strategies based on champion / challenger testing strategy outcomes
- Applying advanced analytical techniques to make specific recommendations

Principa
www.principa.co.za

Data Warehouse, Data Lake: schema-on-write and schema-on-read



Data warehouse Architect

Business intelligence is a term commonly **associated with data warehousing**. If BI is the front-end, data warehousing system is the backend, or the infrastructure for achieving business intelligence. As such, we will first discuss BI in the context of using a data warehouse infrastructure.

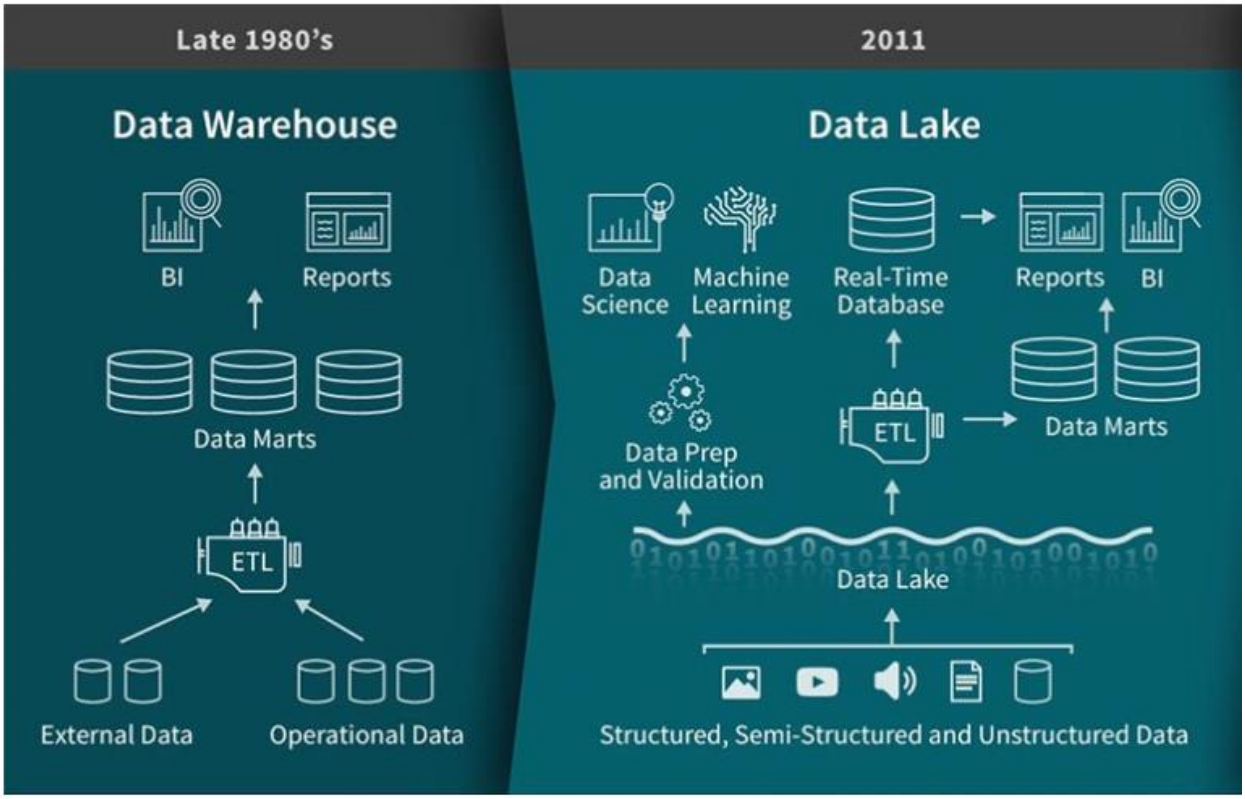
The ultimate purpose of a BI implementation is to **turn operational data into meaningful information**. As raw data is scattered from different operational databases, which are designed and optimized for applications to run rather than for analysis purposes. Sometimes in order to get one data field, you would have to need do ten joins! And here it goes. People come up with a solution called **a central data storage — data warehouse**.

Data warehousing solution emerged in the 1980s, which is optimized for providing information or insights. The data warehouse is a destination database, which centralizes an enterprise's data from all of its source systems. It is a relational database because we can join data from different tables using the joint field as presented in the physical data model. The data base schema defines relations between different tables. Typical SQL databases include mysql and postgresql.

As mentioned, **the data sources connect into the data warehouse through a ETL process, known as Extract, Transform, and Load**. A data warehouse follows the **schema- on- write pattern**, where the design fits the answers to the expected questions. Put differently, **a data warehouse collects data coming from primary applications with predetermined structure and schema**. In Data warehouse architecture, when we move data from a database A to database B, we need to have some information beforehand about the structure of database B and how to adapt the data of database A to fit the structure of data B, for instance to fit the data type of the database B, etc.

Three-layer Datawarehouse Architecture

Data warehouse engineers can use various architectures to build data warehouse. Common data warehouse architectures are based on layer approaches. The well- known **three- layer architecture** is introduced by Inmon, which includes the following components:

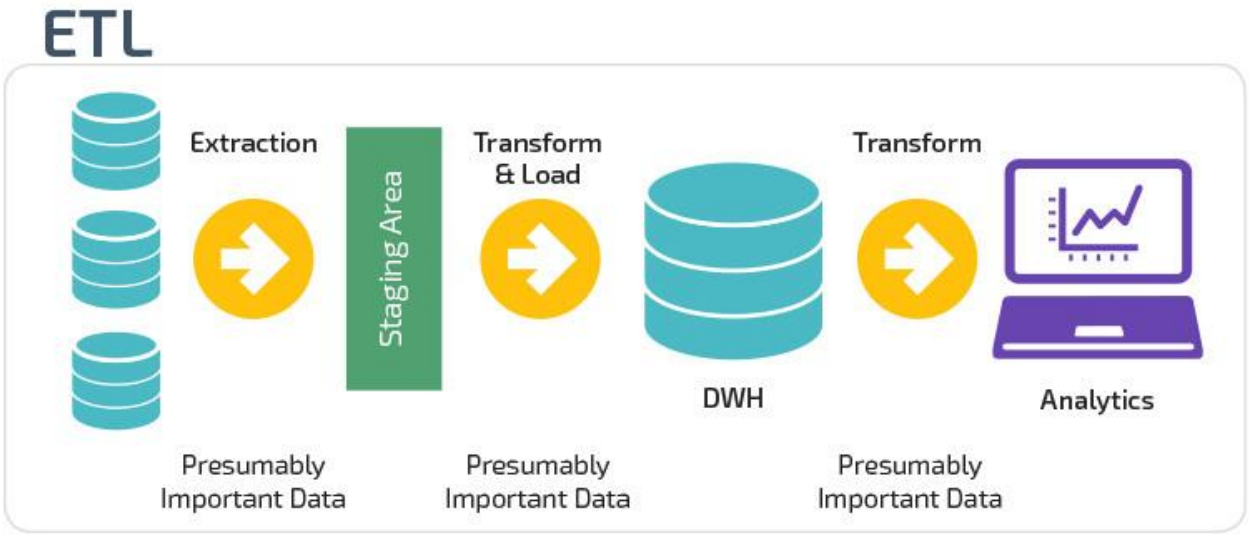


The first layer in line is **Staging area**. This is a data base used to load batch data from source system. Its purpose is to extract the source data from the source systems/primary applications to reduce the work-load on operational system. **Staging Area consists of tables that mirror the structures of the source systems**, which **include all the tables and columns of the sources, including the primary key**. **No functional business rules is applied here**. However, some hard technical business rules are enforced at this stage, for instance, data type matching (string length or unicode characters). This technical business rules do not change the meaning of data, but rather only the way data is stored.

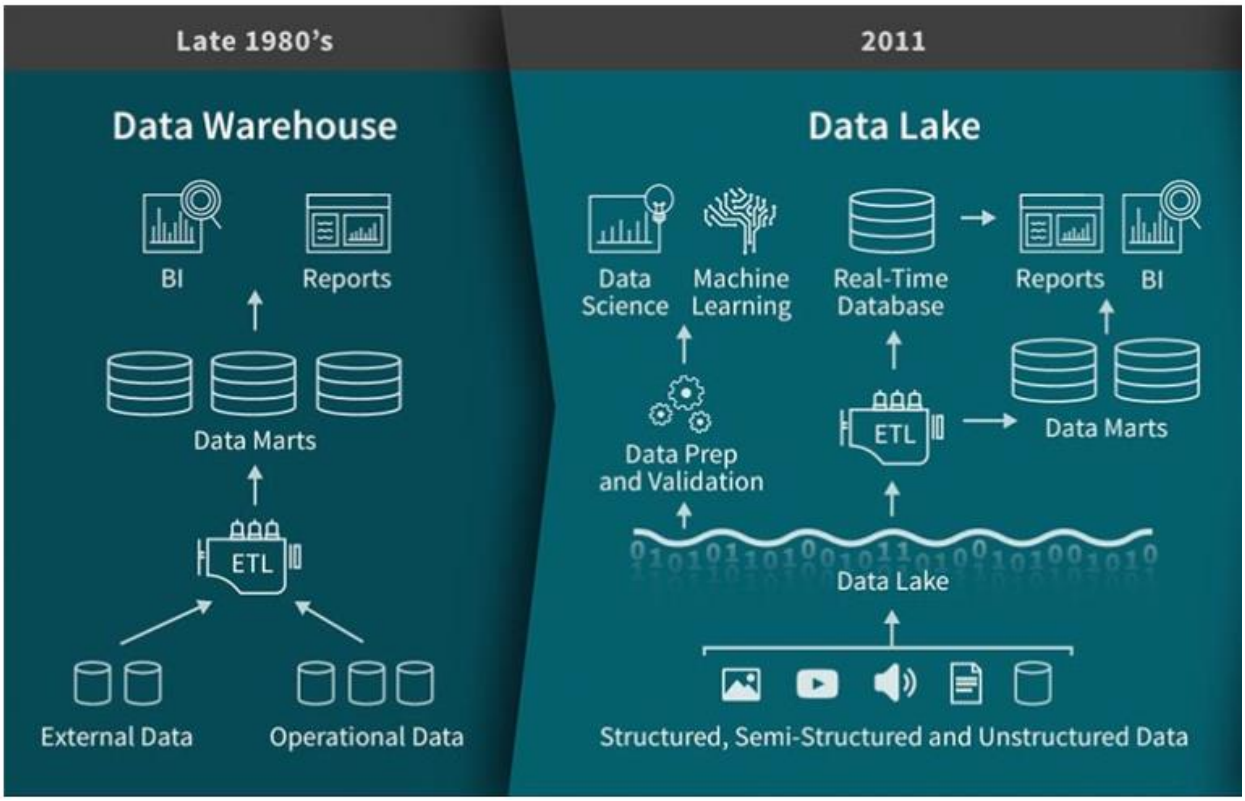
The second layer is the **data warehouse layer**, which is the place where the transformation — **applying functional business rules happens**. These functional business rules modify incoming data to fit the business requirements. The earlier the business rules are implemented a data warehouse architecture, the more dependencies it has on higher layers on top of the data warehouse. In data warehouse modelling, you will often hear about **Dimensional modelling** or **Data Vault modelling**.

Data Vault modelling technique was invented by Dan Linstedlt in the 1990s. It is based on **three basic entities type including the Hubs, The Links and The Satellites**. **The Hubs** are the main pillars of Data Vault Model, which **present the business keys being used by the business users to identify business objects**. Examples of business keys include: Customer identifier, product identifier, employee badge number etc. **The link ties the Hubs together and stores the relationship between different Hubs (business objects)**. **Satellites** is know as **the standard place to store metadata**. Apart from that, **Satellites store all the attributes that describe a business object or a relationship**. They **add the business contexts at a period of time to Hubs and Links**. However, this business context changes over time and the goal of having satellites is also to track those changes and store historical data.

On top of the data warehouse, there is a **dimensional model which builds up the data mart layer** — the layer used to **present the data to the end-user**. I think it makes more sense to call it information mart rather than data mart. The key concepts of dimensional modelling include **fact entities and dimension entities**, which is a standard technique for building the data mart since it is very straightforward for end-users to understand. In dimensional modelling, you will often hear about **star schema** — meaning **one fact table referencing any number of dimension tables. Fact table represents what happens, orders, amounts transactions, meanwhile, dimension entities contain different attributes/fields that constitute the facts**.



Data Lake architecture



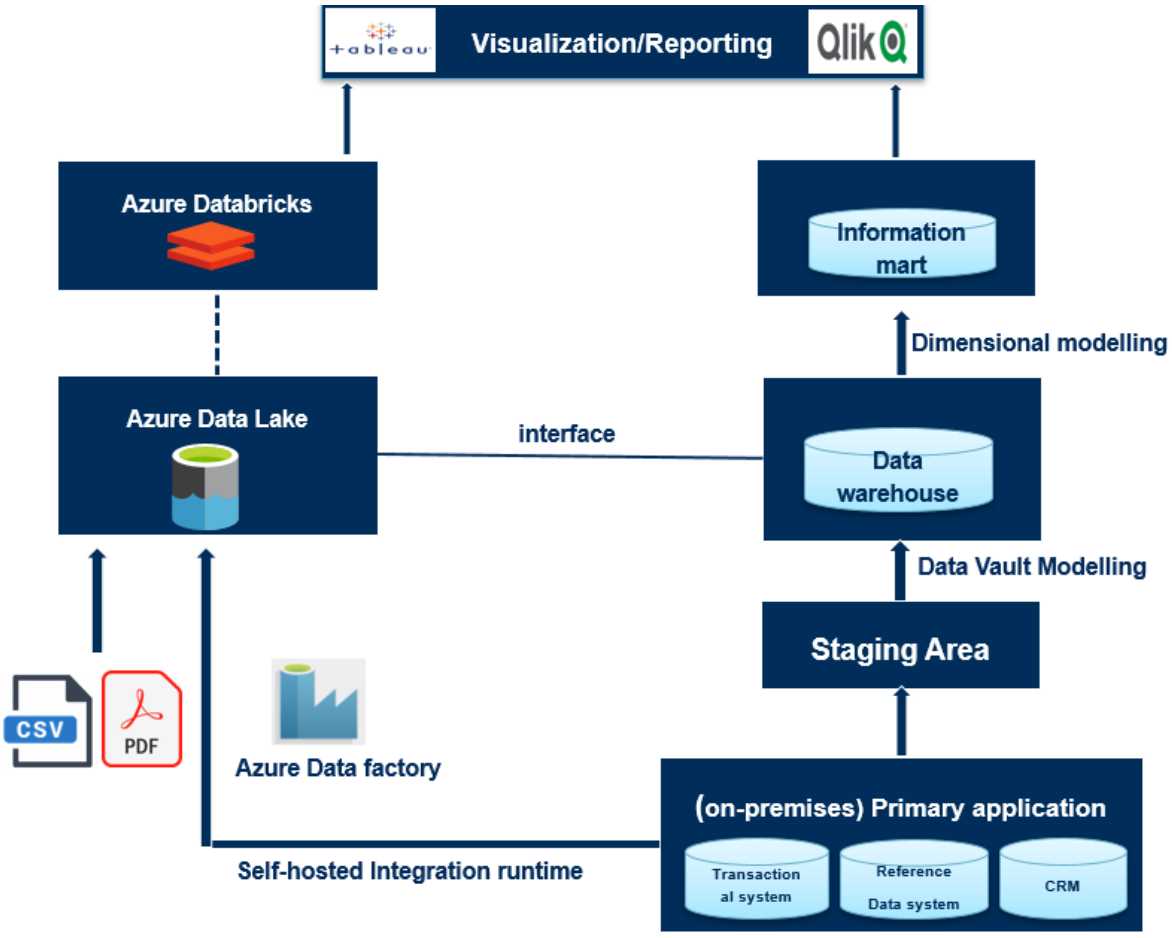
Because data that goes into data warehouses needs to go through a strict governance process before it gets stored, adding new data elements to a data warehouse means changing the design, implementing or refactoring structured storage for the data and the corresponding **ETL** to load the data. **With a massive amount of data, this process could require significant time and resources**. This is where a data lake concept comes into the picture and becomes a game-changer in big data management.

The concept of **data lake** emerges in the 2010s, which, in a simple language, is the idea that **all enterprise's structured, unstructured and semi-structured data can and should be stored in the same place**. Apache Hadoop is an example of data

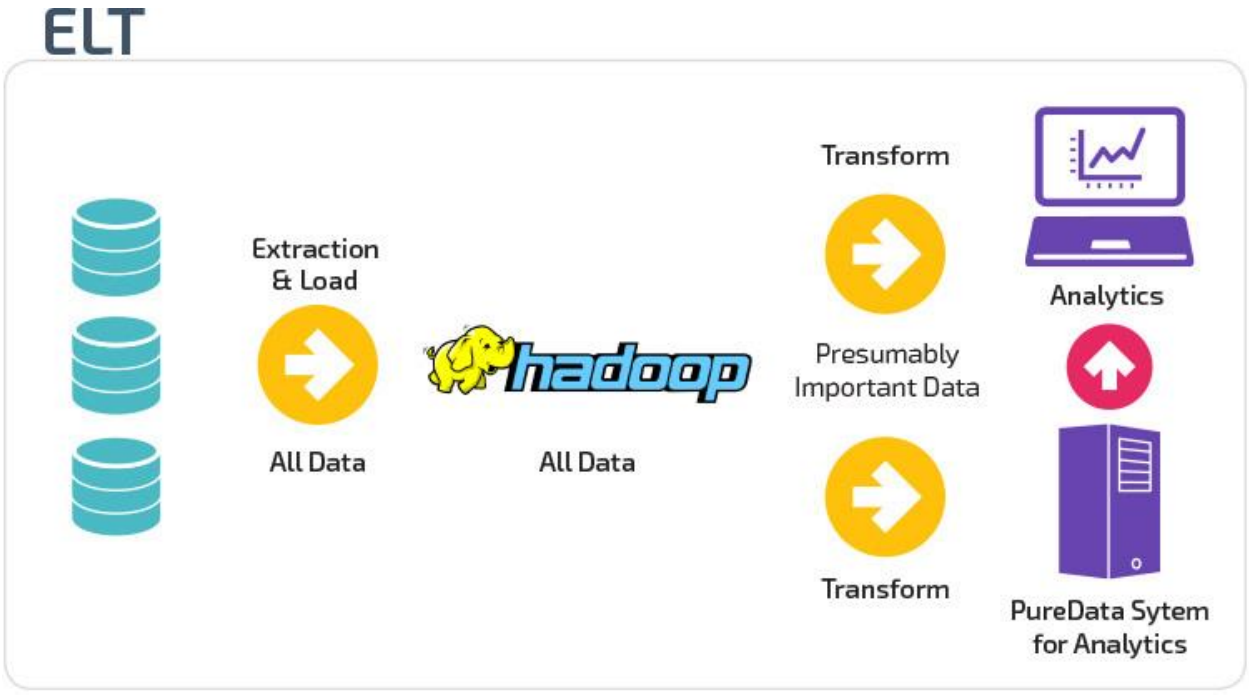
infrastructure that allows to store and process massive amounts of data, both structured and unstructured; which enables the Data Lake architecture.

A data lake is a collection of storage instances of various data assets additional to the originating data sources. These assets are stored in a near-exact, or even exact, copy of the source format. The purpose of a data lake is to present an unrefined view of data to only the most highly skilled analysts, to help them explore their data refinement and analysis techniques independent of any of the system-of-record compromises that may exist in a traditional analytic data store (such as a data mart or data warehouse).

— Gartner



Data lake has schema on read approach. It stores raw data and is set up in a way that does not require defining the data structure and schema in the first place. Put differently, when we move data to data lake, we just bring it in without any gate keeping rules and when we need to read the data, we apply the rule to the code that reads the data rather than configuring the structure of data ahead of time. Instead of the typical Extract, Transform and Load in data warehousing, in the world of data lake, the process is Extract, Load and Transform. Data Lake is utilized for cost efficiency and exploration purpose. As such, **a Data Lake architecture enables business to gain insights not only from the processed and governed data but also from raw data that was not available for analysis before**. From that, raw data exploration can potentially trigger business questions. However, the biggest concern with data lake is that, without appropriate governance, data lakes can quickly turn into unmanageable data swamps. Put it differently, without knowing how the water is in a lake, who would want to go swim in it? Business users can't not utilize the data lake if they don't trust the data quality of that lake.



Recently the trend of companies that want to benefit from a data lake architecture in a more conservative way has emerged. These companies are stepping away from the ungoverned "free-entry" approach and instead developing a more governed data lake.

The **data lake can contain two environments**: **an exploration/development** and **production environment**. **Data will be explored, cleansed, transformed in order to build machine learning model, build functions and other analytics purposes**. **Data such as metrics, functions that have been generated by the transformation process will be stored in the production part of data lake**.

Another trend is, rather than pouring all raw data into the lake, the governed data lake only allows 'verified' data to get into it. Essentially, a governed data lake architecture does not restrict the types of data that are stored in it, meaning that governed data lakes still comprise multiple data types including unstructured and semi-structured data like XML, JSON, CSV. However, the key is to make sure that no data is stored in the lake without being described and documented in business glossary, which will give some confidence to the users about the quality and meaning of data.

To provide this layer of governance, a business glossary tool has to be in place to document the meaning of the data. More importantly, there needs to be a governance process around this — which is all about roles and responsibilities, for instance, who owns the data, who defines it, who will be responsible for any data quality issues. Going for this approach will be time-consuming because defining data itself can be a long process since it involves people from different disciplines across an enterprise.

On-premise and cloud service

In the old days, companies usually rely on self-host-data-centre. During data processing, a huge amount of data has to be processed and data processing runs on a cluster of machines at the same time rather than on one computer. This on-premise storage set-up centre requires lots of serves and big computers and enough processing power for peak moments. It also means that,

in quieter times, much of the processing power remains unused. Therehore, on-premise systems have huge up-front installation cost but at the same time, their capacity can not always be fully used. Not only that, maintaining this requires a whole dedicated application team to take care of these primary systems, to fix production issues and to reduce the downtime.

These huge disadvantages of on-premise storage has made the idea of using cloud becomes so appealing. Many companies have moved to the cloud as a way of cost optimization. The rise of building data infrustructure on the clouds, for instance using Azure and AWS has completely changed an enterprise's big data capabilities. Instead of maintaining CPUs and storage in their data center, companies only need to rent the cloud resources, for example, the storage they need at the time they need it. The cloud providers take care of maintaining the underlying infrastructure. When cloud-service users are done using the resouces, they can give them back and are billed only for what they use. This makes it very easy to scale since if things are slowing down, companies can assign some more virtual machines to do data processing for instance. With cloud storage, enterprises don't have to invest in new machines, infrastructure, or replace aging servers. Another reason for using cloud storage is database reliability. In the worse case, for example, a fire can happen at your data centre. To be safe, you need to replicate your data to different geographical locations. This brings a bunch of logistic problems. On these needs, companies specilizing in this kind of service were born. We call them "cloud-service providers". There are three big players on cloud-service provider market. First and foremost, Amazon Web service (AWS) is the market leader, who takes up 32% of market share in 2018. Second position on the run belongs to Microsoft Azure that owns 17% of the market share and third in popularity is Google Cloud Services who owns 10% of market share.

What do these cloud-service providers have to offer?

There are a bunch of cloud services that all have different use cases. The most simple service is storage, which allows cloud-service users to upload all type of files on the cloud. Storage services are typically cheap since they don't provide much functionality other than storing the files reliably. AWS hosts S3 as a storage service. Azure has Blob storage and google has Cloud Storage. The second service is computation service which allows you to perform computation on the cloud. Essentially, you can start up a virtual machine and use it as you wish. Computation service is often used to host web servers and performing calculations, for example. AWS provides EC2 as

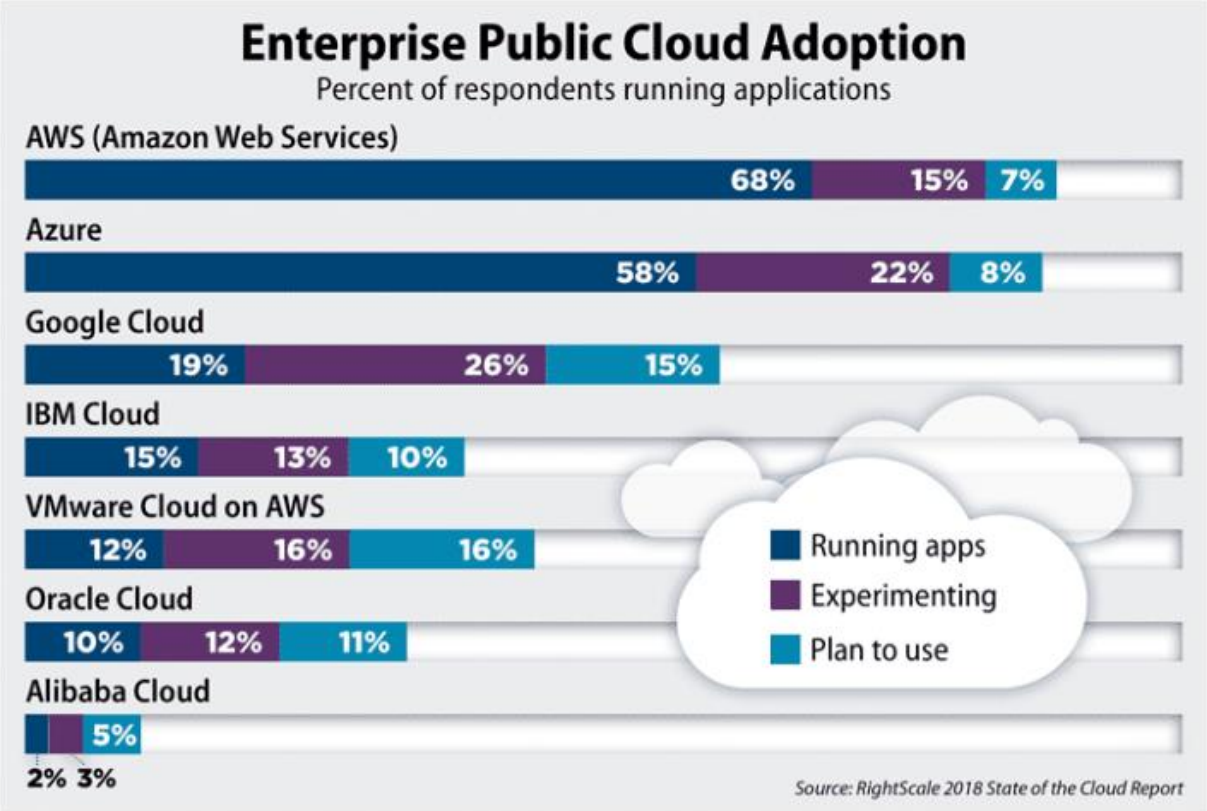
computation service, Microsoft Azure uses Virtual machines and Google has Compute Engine. Cloud-service providers also host databases. For instance, companies can start exploring the cloud by moving their existing applications to virtual machines that run in Azure. Another use cases of cloud service is Artificial Intelligence and Machine Learning. Azure for example provides Azure Machine Learning service which is a cloud-based environment you can use to develop, train, test and deploy machine learning models.

Can a Data Lake replace Data warehouse for BI?

It is no doubt that a Data Lake offers several benefits over data warehouse, particularly for dealing with big data and cost efficiency. However, without knowing about the data quality in data lakes, business users can't not trust in that data. Lots of data scientists often don't realize that getting the raw data from multiple sources to be ready — clean and high quality for modelling usually takes 80% of the time. Meanwhile, the architecture and discipline of Data Warehouses that provide governed and good quality data can't be ignored. There is still a strong and permanent demand for a core set of KPIs that define how a business is doing, which are also needed for reporting, especially regulatory reporting purposes. Such information demands highly governed information.

At the end of the day, the most crucial question about what companies want to achieve in their business strategy with data and technology has to be crystal clear. For example, if a company, in their business and data strategy, does not plan to be involved in data science, there is no point of investing in Data Science/Analytics platform that allows AI and Machine Learning to happen. Or companies can govern all its data in a business glossary platform and helicopter them into a Data Lake, but if they don't know why, I doubt if that is going to create value.

Data Lake	Data warehouse
Schema on read	Schema on write.
Can store structure, unstructured and semi-structured data	Mainly store structure data
Is optimized for cost efficiency	Is optimized for reporting purpose.
Store raw data	Store processed and governed data
ELT process	ETL process



Inside the Data Warehouse and Data Lake

For **a firm that's looking to analyze large but structured data sets, a data warehouse is a good option**. In fact, **if the company is only interested in descriptive analytics — the process of merely summarizing the data one has — a data warehouse may be all it needs**.

Let's say, for example, company leaders want to look at sales figures across a particular time period, the number of inquiries about a product, or the view counts on various marketing videos. A data warehouse would be perfect for those applications because all of the associated figures are stored in the form of structured data.

But **for most companies embarking on big data initiatives, structured data is only part of the story**. Each year, businesses generate a staggering quantity of unstructured data. In fact, 451 Research in conjunction with Western Digital found that 63 percent of enterprises and service providers are keeping at least 25 petabytes of unstructured data. For those firms, **data lakes are attractive options because of their ability to store vast quantities of such data**.

What's more, **data lakes allow analysts to go beyond descriptive analytics and into the exciting — and highly rewarding — domain of predictive or prescriptive analytics**. Predictive analysis is the practice of using existing data to predict future trends relevant to one's business, such as next year's revenue.

Prescriptive analytics goes a big step further, using artificial intelligence technologies to make recommendations in response to predictions. For both predictive and prescriptive analytics, a data lake is a must. Often, leaders manage data lakes using software like Apache Hadoop, a popular ecosystem of analytics tools.

เพิ่มเติม

- **การ์ทเนอร์ (Gartner, Inc.) คือ บริษัทวิจัยและให้คำปรึกษาด้านเทคโนโลยีสารสนเทศ**ชั้นนำของโลก บริษัทฯให้ข้อมูลเชิงลึกที่เกี่ยวข้องกับเทคโนโลยีให้แก่ลูกค้าเพื่อใช้ประกอบการตัดสินใจ การ์ทเนอร์มีฐานลูกค้าตั้งแต่กลุ่มซีไอโอ ผู้บริหารงานไอทีระดับสูงทั้งในองค์กรภาครัฐและเอกชน รวมถึงผู้นำทางธุรกิจในสายงานที่เกี่ยวข้องกับเทคโนโลยีและโทรคมนาคม บริษัทผู้ให้บริการทางด้านเทคโนโลยีสารสนเทศชั้นนำต่างๆ ไปจนถึงหน่วยงานที่ลงทุนด้านเทคโนโลยี การ์ทเนอร์เป็นพันธมิตรที่สำคัญของลูกค้ามากกว่า 10,000 องค์กรทั่วโลก การ์ทเนอร์ทำงานร่วมลูกค้าตั้งแต่การทำวิจัย วิเคราะห์และตีความข้อมูลทางธุรกิจให้สอดคล้องกับบริบทของลูกค้าในแต่ละตำแหน่ง โดยใช้งานวิจัยจากการ์ทเนอร์ ร่วมกับ Gartner Executive Program และ Gartner Consulting รวมถึง Gartner Event ต่าง ๆ การ์ทเนอร์ก่อตั้งขึ้นในปี 2522 โดยมีสำนักงานใหญ่อยู่ที่ แสตมป์ฟอร์ด รัฐคอนเนตทิคัต สหรัฐอเมริกา ซึ่งมีพนักงานกว่า 8,300 คน รวมไปถึงนักวิเคราะห์และที่ปรึกษากว่า 1,800 คน และมีฐานลูกค้าอยู่ในกว่า 90 ประเทศ สำหรับข้อมูลเพิ่มเติม เยี่ยมชมได้ที่ www.gartner.com

Ref.

- https: / / towardsdatascience. com/ implementing- a- data- lake- architecture- for- business-intelligence-f2c99551db1a
- https://panoply.io/data-warehouse-guide/
- https://www.techtalkthai.com/gartner-in-bangkok/
- https://medium.com/@thanachart.rit/data-lake-%E0%B8%84%E0%B8%B7%E0%B8%AD%E0%B8%AD%E0%B8%B0%E0%B9%84%E0%B8%A3-%E0%B9%80%E0%B8%84%E0%B9%89%E0%B8%B2%E0%B8%9E%E0%B8%B9%E0%B8%94%E0%B8%96%E0%B8%B6%E0%B8%87%E0%B8%81%E0%B8%B1%E0%B8%99%E0%B8%AD%E0%B8%A2%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B9%84%E0%B8%A3-%E0%B9%83%E0%B8%99-google-cloud-next18-5e6fba897b11
- https://medium.com/readwrite/data-lake-vs-data-warehouse-which-is-the-best-data-architecture-ebda7192f99f
- https://thanachart.org/2018/03/27/%E0%B8%88%E0%B8%B0%E0%B8%97%E0%B8%B3-big-data-%E0%B8%95%E0%B9%89%E0%B8%AD%E0%B8%87%E0%B9%80%E0%B8%A3%E0%B8%B4%E0%B9%88%E0%B8%A1%E0%B8%95%E0%B9%89%E0%B8%99%E0%B8%97%E0%B8%B5%E0%B9%88%E0%B8%97%E0%B8%B3/
- https://thanachart.org/2017/11/24/%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%97%E0%B8%B3%E0%B9%82%E0%B8%84%E0%B8%A3%E0%B8%87%E0%B8%81%E0%B8%B2%E0%B8%A3-big-data-%E0%B8%AD%E0%B8%A2%E0%B9%88%E0%B8%B2%E0%B8%87%E0%B8%A3%E0%B8%A7%E0%B8%94%E0%B9%80/
- https://thanachart.org/2018/01/18/big-data-%E0%B8%95%E0%B9%89%E0%B8%AD%E0%B8%87%E0%B9%80%E0%B8%A3%E0%B8%B4%E0%B9%88%E0%B8%A1%E0%B8%95%E0%B9%89%E0%B8%99%E0%B8%88%E0%B8%B2%E0%B8%81%E0%B8%81%E0%B8%B2%E0%B8%A3%E0%B8%A7%E0%B8%B4%E0%B9%80/
- https://www.coraline.co.th/single-post/Data-Lake-Data-Governance-Data-Analytics-Data-Cleansing