Big Data Fundamental

Presenter

in /arifmazumder

Mohammed Arif, PhD

Lead Data Scientist

Big Data | Machine Learning | AI

Microsoft CERTIFIED AZURE DATA SCIENTIST ASSOCIATE

Microsoft CERTIFIED AZURE DATA ENGINEER ASSOCIATE

CCA cloudera SPARK & HADOOP DEVELOPER

Mohammed Arif has more than fifteen (15) years of working experience in Information Communication and Technology (ICT) industry. The highlights of his career are more than six (7) years of holding various senior management and/or C-Level and had five (5) years of international ICT consultancy exposure in various countries (APAC and Australia), specially on Big Data, Data Engineering, Machine Learning and AI arena.

He is also Certified Trainer for Microsoft & Cloudera.

MICROSOFT CERTIFIED TRAINER Microsoft MCT Trainer 2020 - 2021

Agenda

[Day 1]

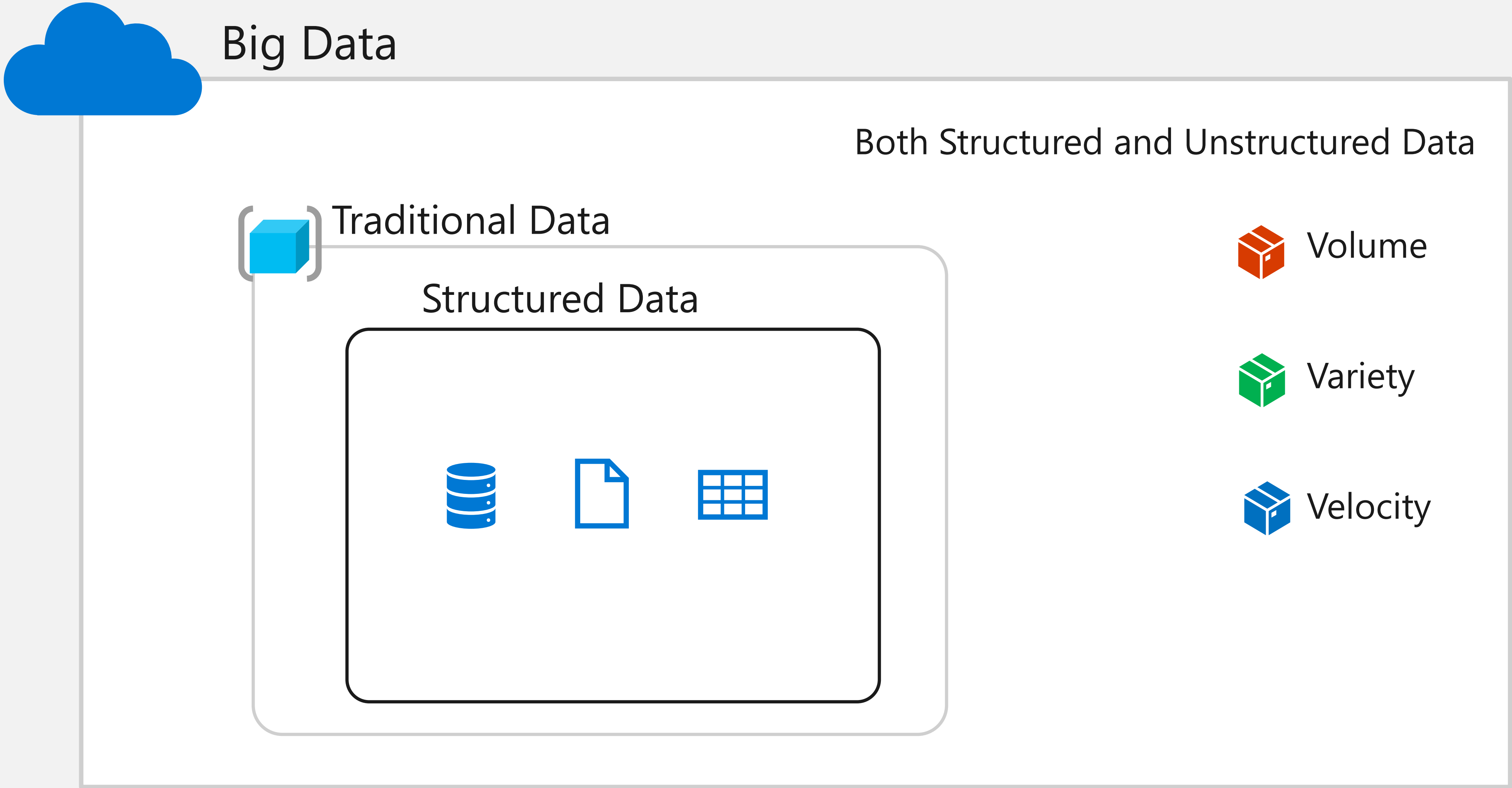
- Data vs Big Data
- Big Data Characteristics
- Big Data Reference Architecture
- Big Data Ecosystem Components
- Analysis vs Analytics
- Data Analysis (Hands-on)
- Big Data – Career Path

Resource Link

http://arif.works/bdf/

Data vs Big Data

Data



Big Data

Both Structured and Unstructured Data

Traditional Data

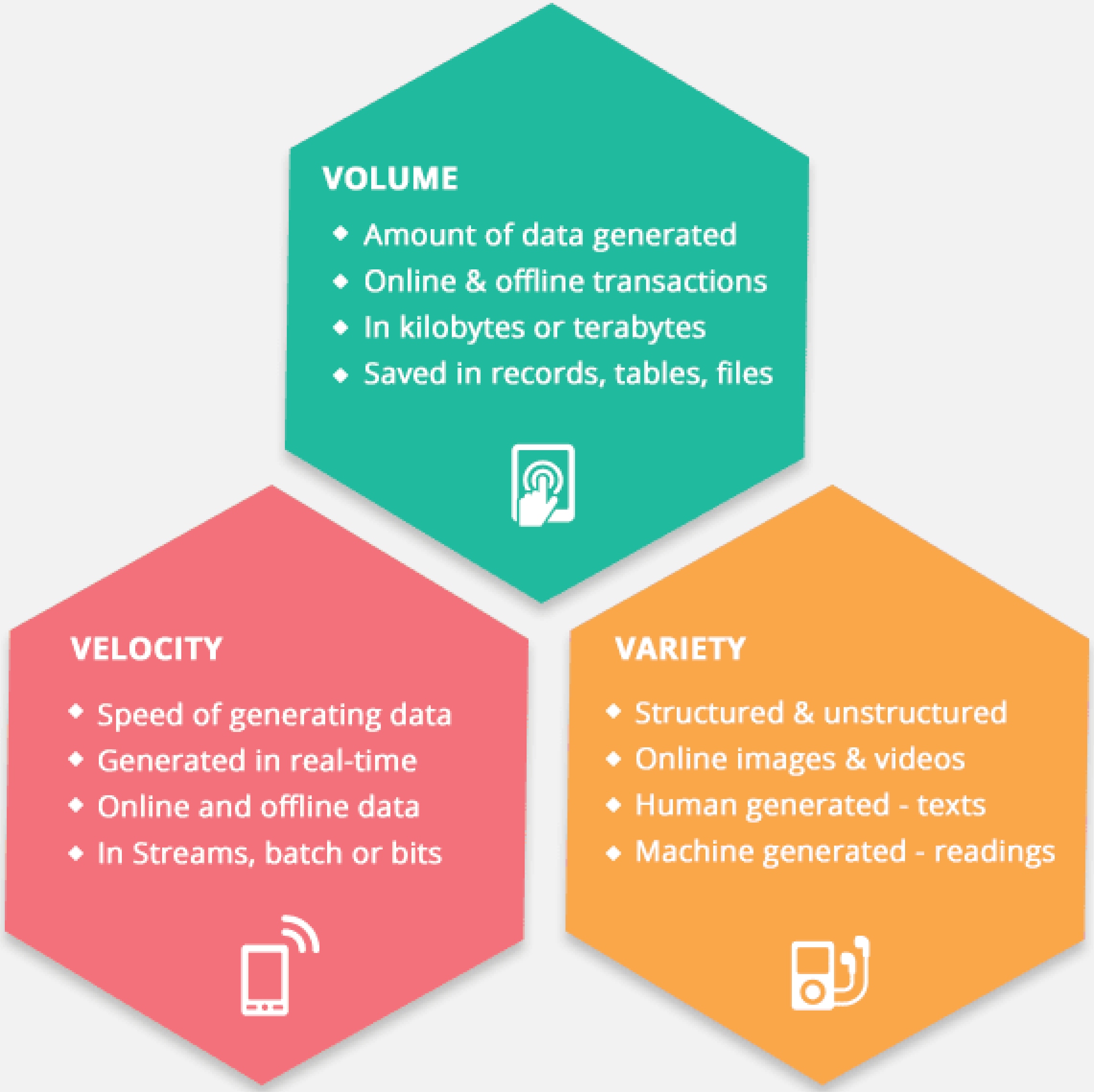
Structured Data

Volume

Variety

Velocity

Big Data Characteristics



VOLUME

- Amount of data generated
- Online & offline transactions
- In kilobytes or terabytes
- Saved in records, tables, files

VELOCITY

- Speed of generating data
- Generated in real-time
- Online and offline data
- In Streams, batch or bits

VARIETY

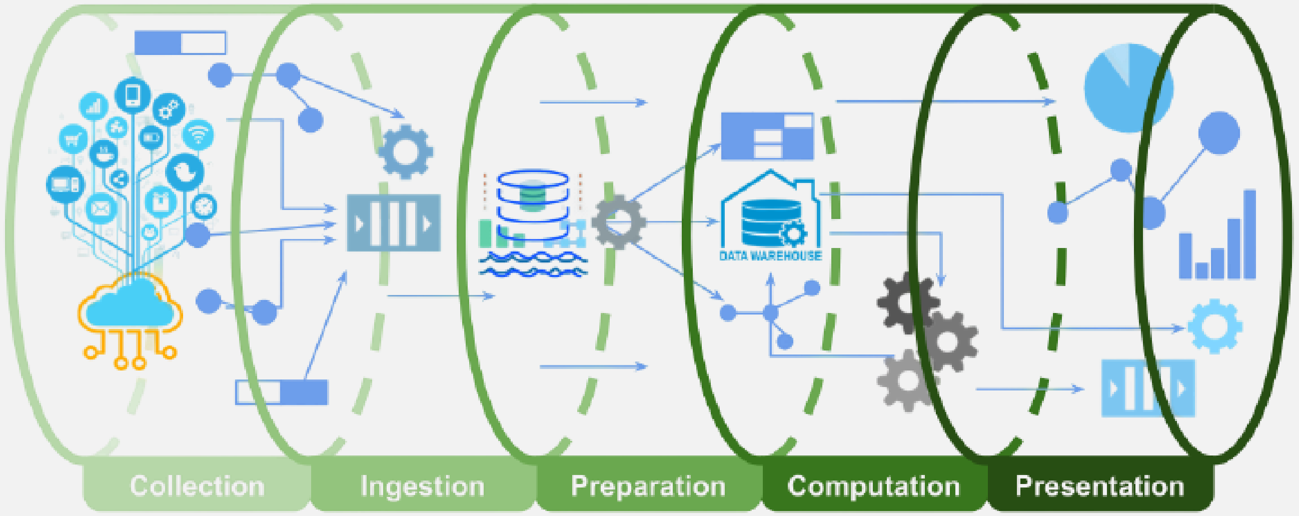
- Structured & unstructured
- Online images & videos
- Human generated - texts
- Machine generated - readings

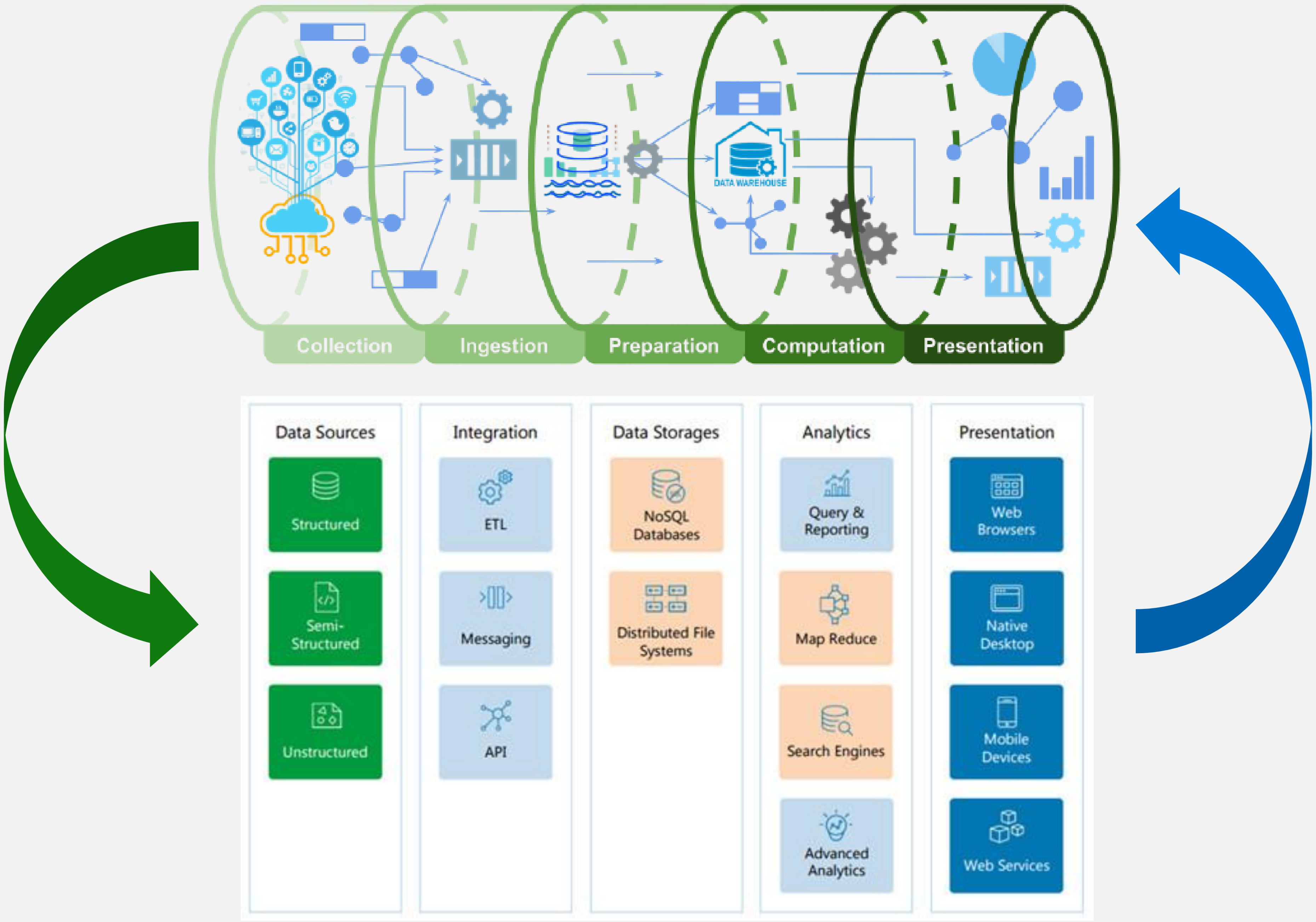
Big Data Reference Architecture

In summary,

Generally Big Data Architecture Data Pipeline has five stages:

- Collection
- Ingestion
- Preparation
- Computation
- Presentation

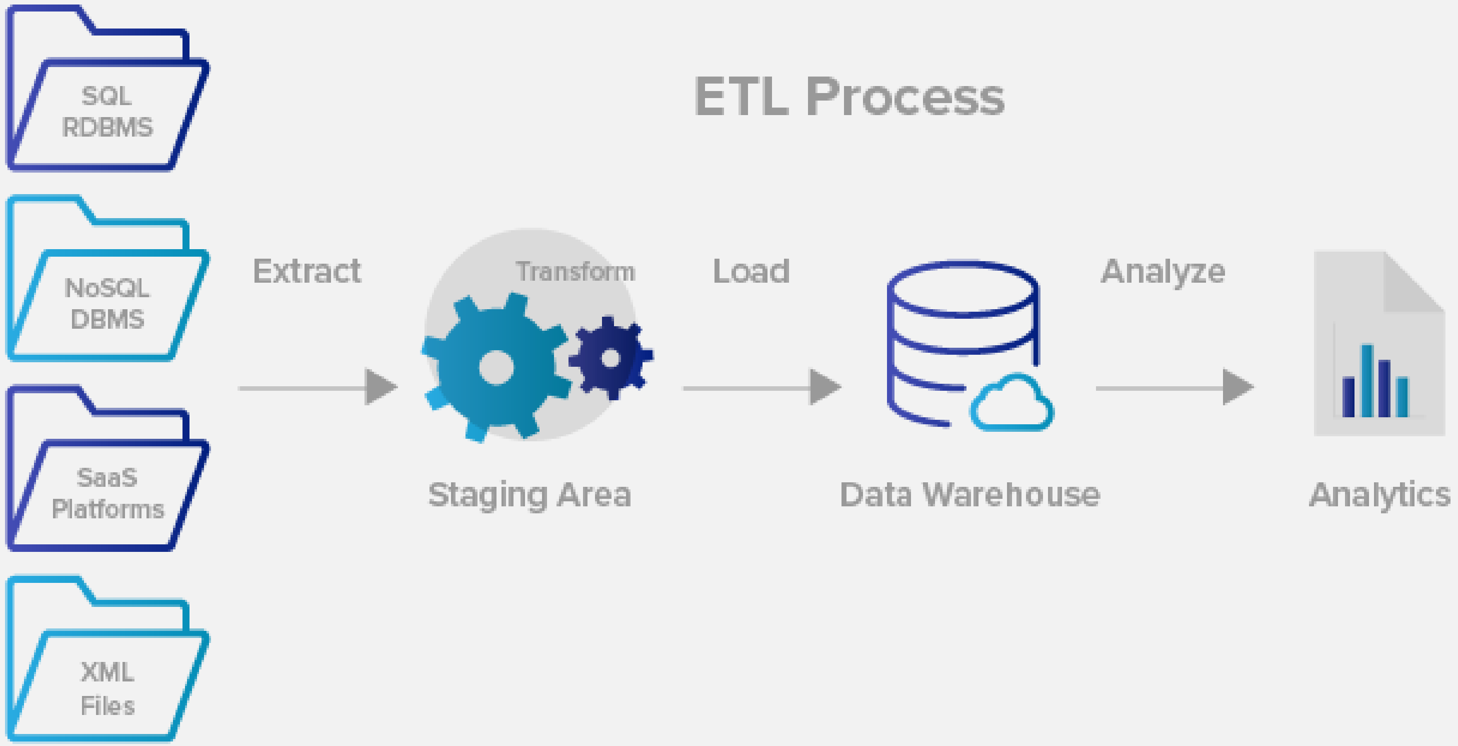




Data Sources	Integration	Data Storages	Analytics	Presentation
Structured	ETL	NoSQL Databases	Query & Reporting	Web Browsers
Semi-Structured	Messaging	Distributed File Systems	Map Reduce	Native Desktop
Unstructured	API		Search Engines	Mobile Devices
			Advanced Analytics	Web Services

SIMPLIFIED ACADEMY
learn • earn • grow

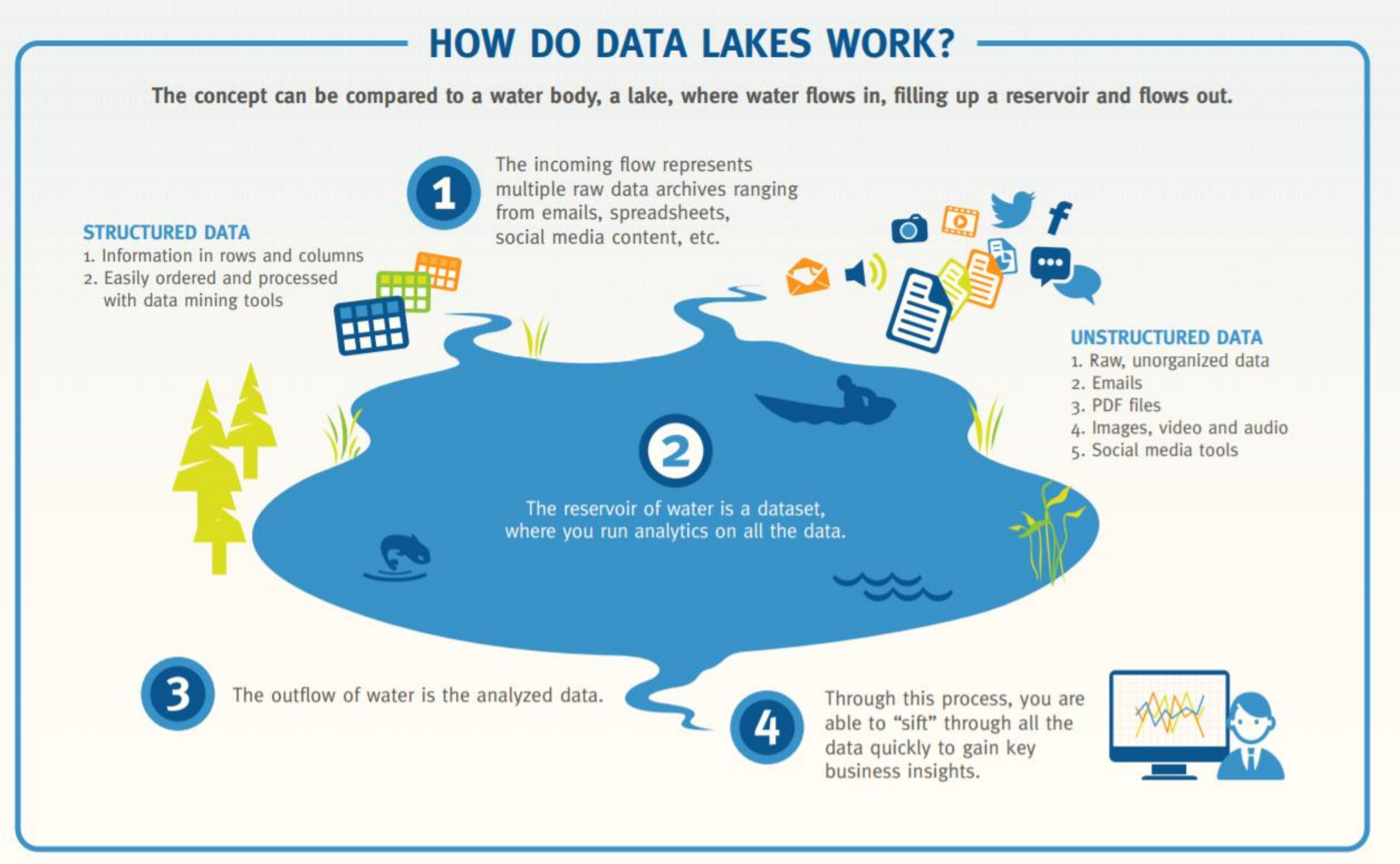
ETL (Extract, Transform & Load)



Data Lake

Data lake is one place to put all the data enterprises may want to use, including structured and unstructured data.

Data Lake



HOW DO DATA LAKES WORK?

The concept can be compared to a water body, a lake, where water flows in, filling up a reservoir and flows out.

1 The incoming flow represents multiple raw data archives ranging from emails, spreadsheets, social media content, etc.

STRUCTURED DATA

1. Information in rows and columns
2. Easily ordered and processed with data mining tools

UNSTRUCTURED DATA

1. Raw, unorganized data
2. Emails
3. PDF files
4. Images, video and audio
5. Social media tools

2 The reservoir of water is a dataset, where you run analytics on all the data.

3 The outflow of water is the analyzed data.

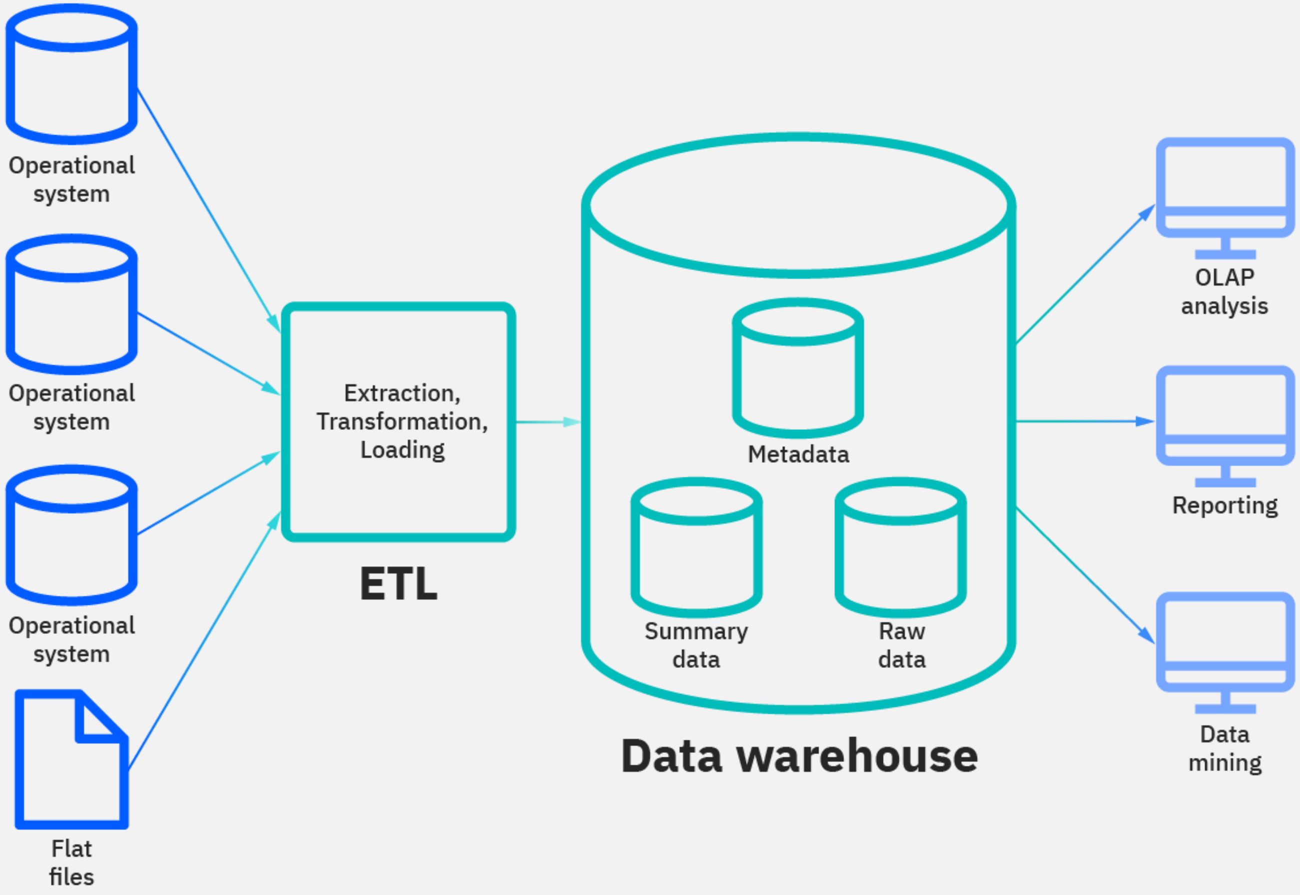
4 Through this process, you are able to "sift" through all the data quickly to gain key business insights.

Data Warehouse

Single Source of Truth.

Structuring all the Best Quality data in one place.

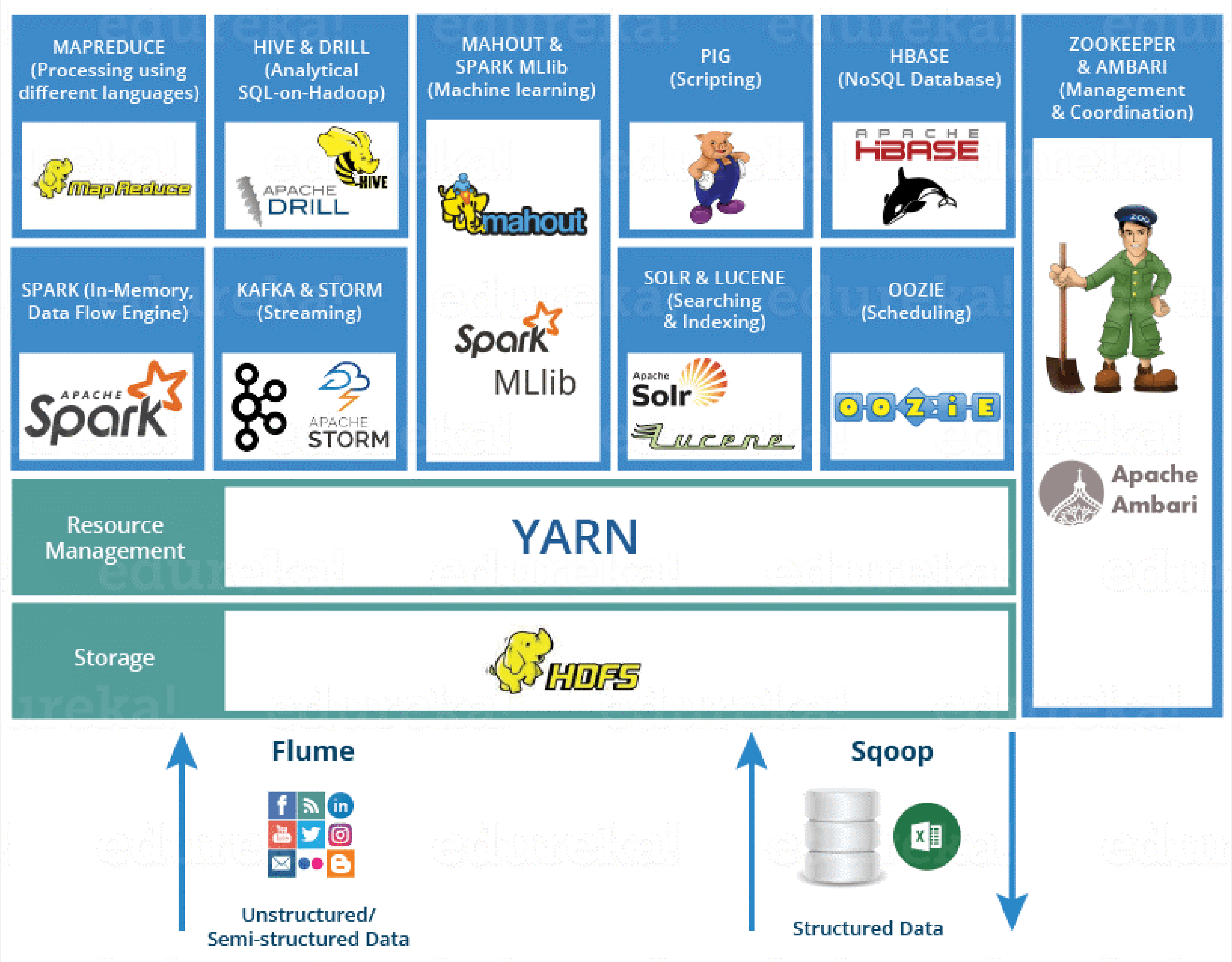
Data Warehouse



Hadoop

Hadoop is a collection of open source programs/procedures/platform relating to Big Data analysis. Being open source, it is freely available for use, reuse and modification (with some restrictions) for anyone who is interested in it. Big Data scientists call Hadoop the 'backbone' of their operations.

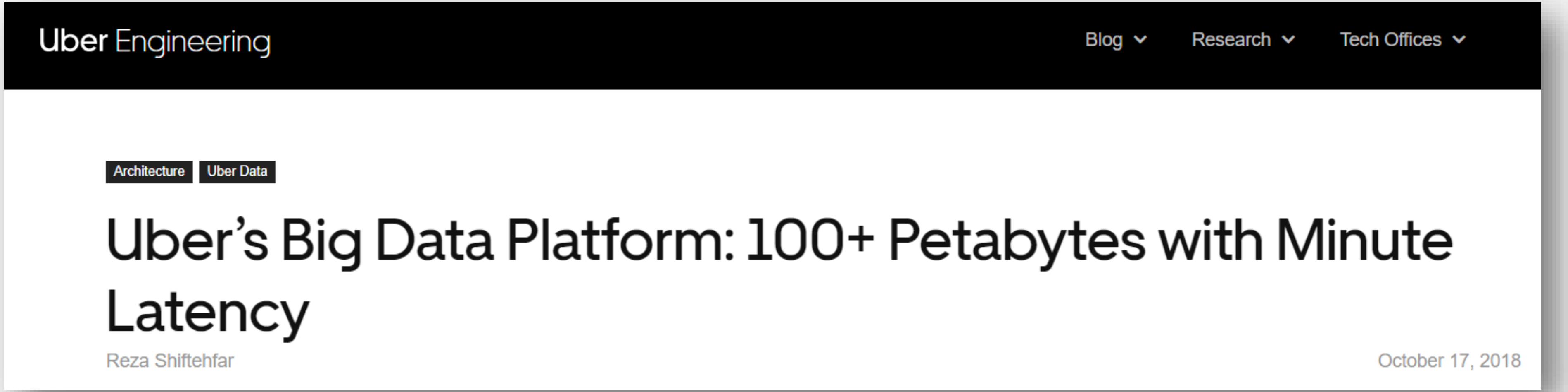
Big Data/Hadoop Eco System Component


MAPREDUCE (Processing using different languages)
MapReduce
HIVE & DRILL (Analytical SQL-on-Hadoop)
HIVE
APACHE DRILL
MAHOUT & SPARK MLlib (Machine learning)
mahout
Spark MLlib
PIG (Scripting)
HBASE (NoSQL Database)
APACHE HBASE
ZOOKEEPER & AMBARI (Management & Coordination)
Apache Ambari
SPARK (In-Memory, Data Flow Engine)
APACHE Spark
KAFKA & STORM (Streaming)
APACHE STORM
SOLR & LUCENE (Searching & Indexing)
Apache Solr
Lucene
OOZIE (Scheduling)
OOZIE
Resource Management
YARN
Storage
HDFS
Flume
Unstructured/ Semi-structured Data
Sqoop
Structured Data

Case Study (Uber)

Transformation Journey towards Big Data Platform.

Please read this article to get more info on how Big Data & its Architecture

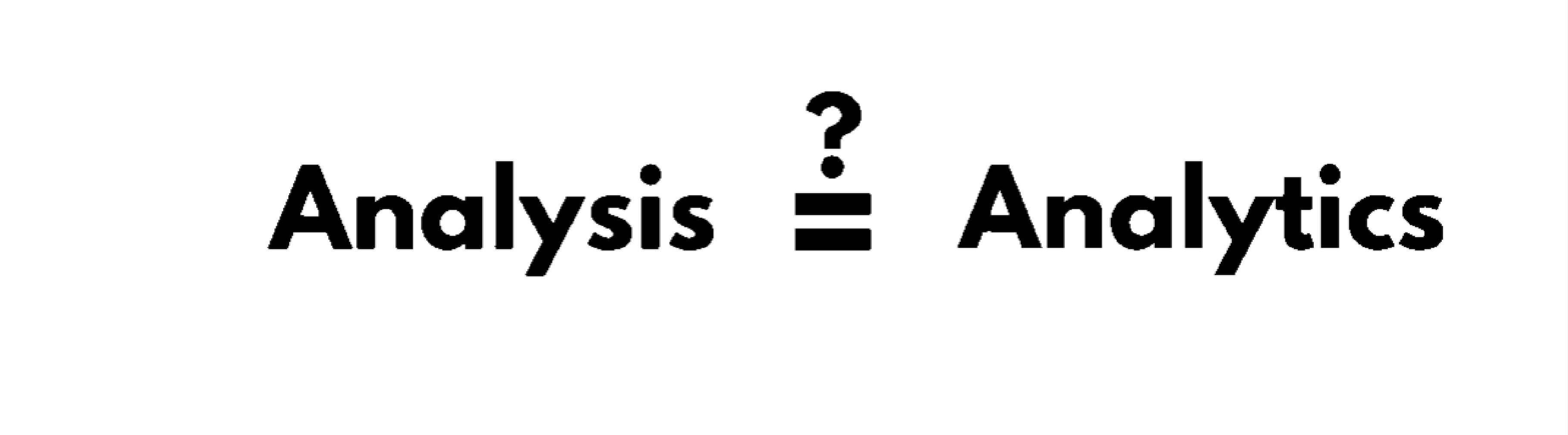


Uber's Big Data Platform: 100+ Petabytes with Minute Latency

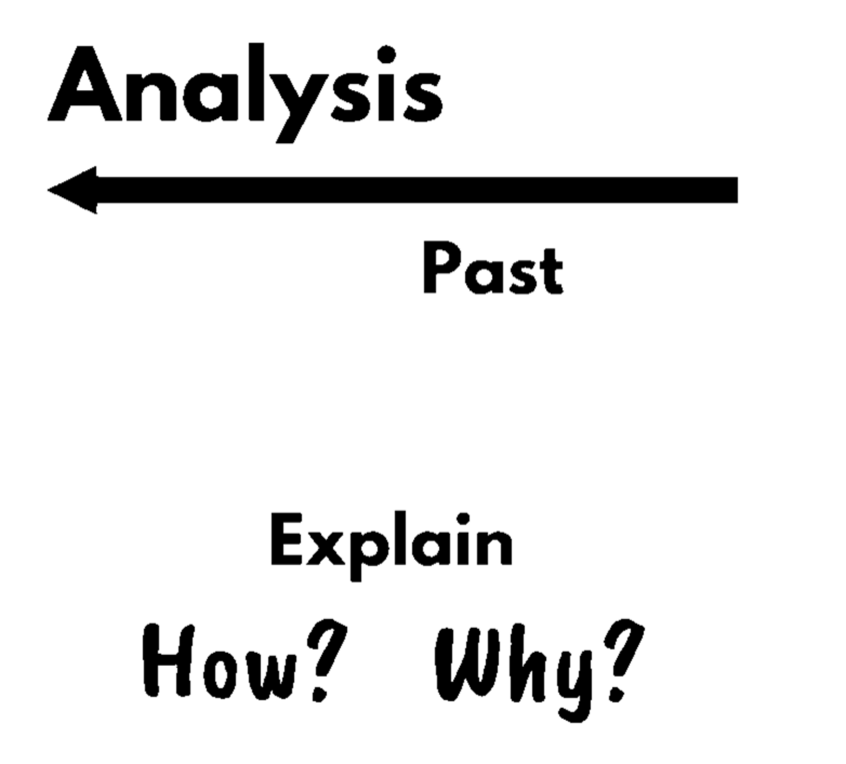
Reza Shiftehfar

October 17, 2018

https://eng.uber.com/uber-big-data-platform/



Analysis $\overset{?}{=}$ Analytics



Analysis

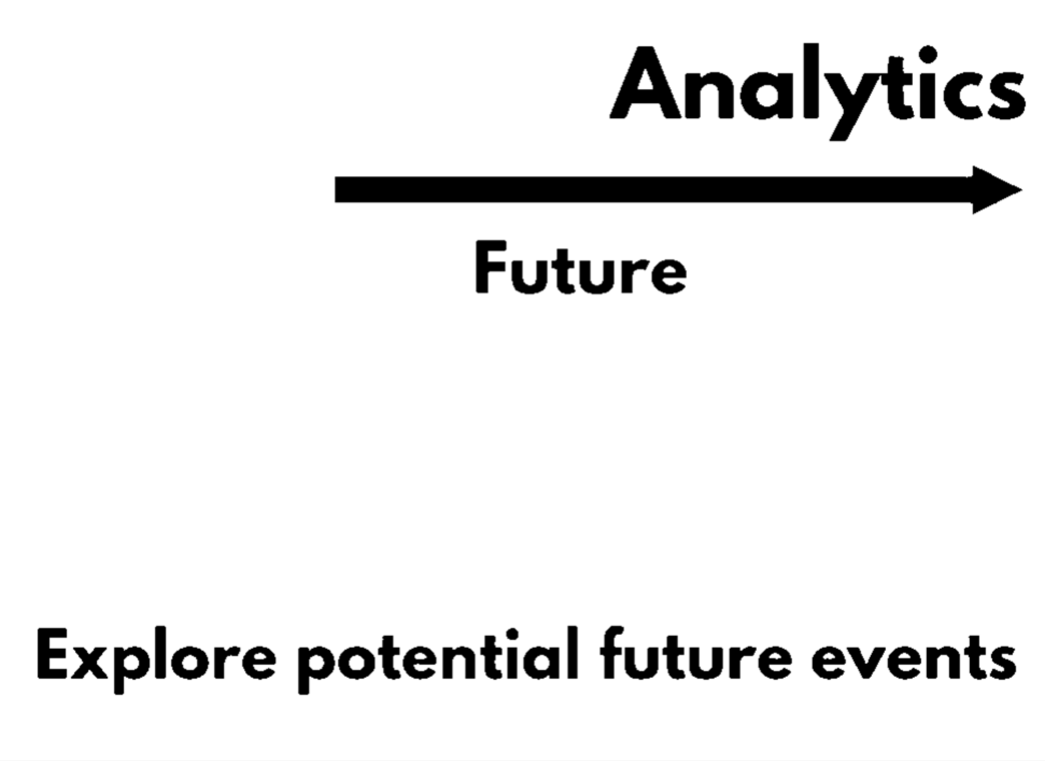
Past

Explain

How? Why?

BI

We can use different tools to explain the previous trends like, PowerBI, Tableau, Qlikview etc.



Analytics

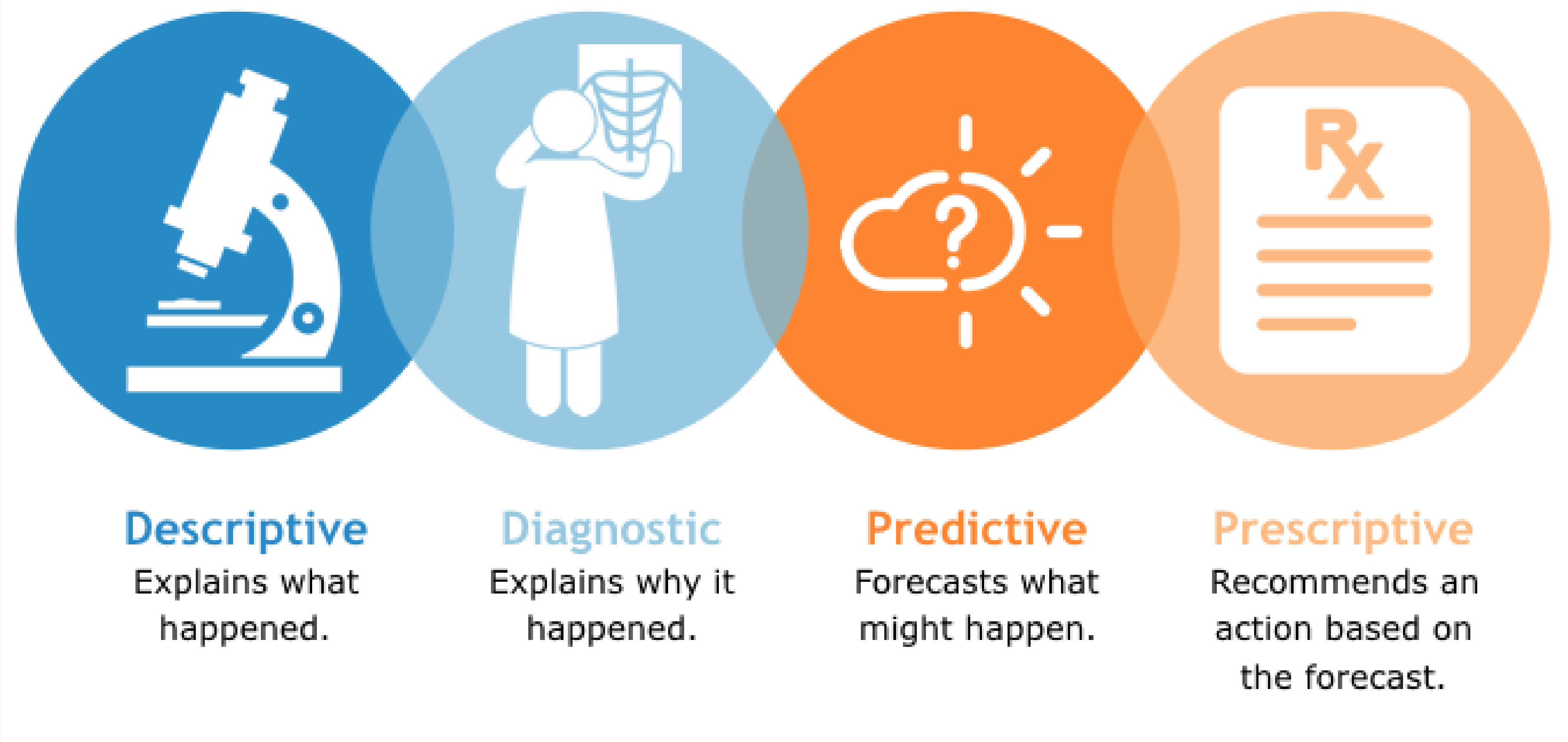
Future

Explore potential future events

ML/AI

We can use different language packages and framework to implement ML/AI model.

Business Analytics



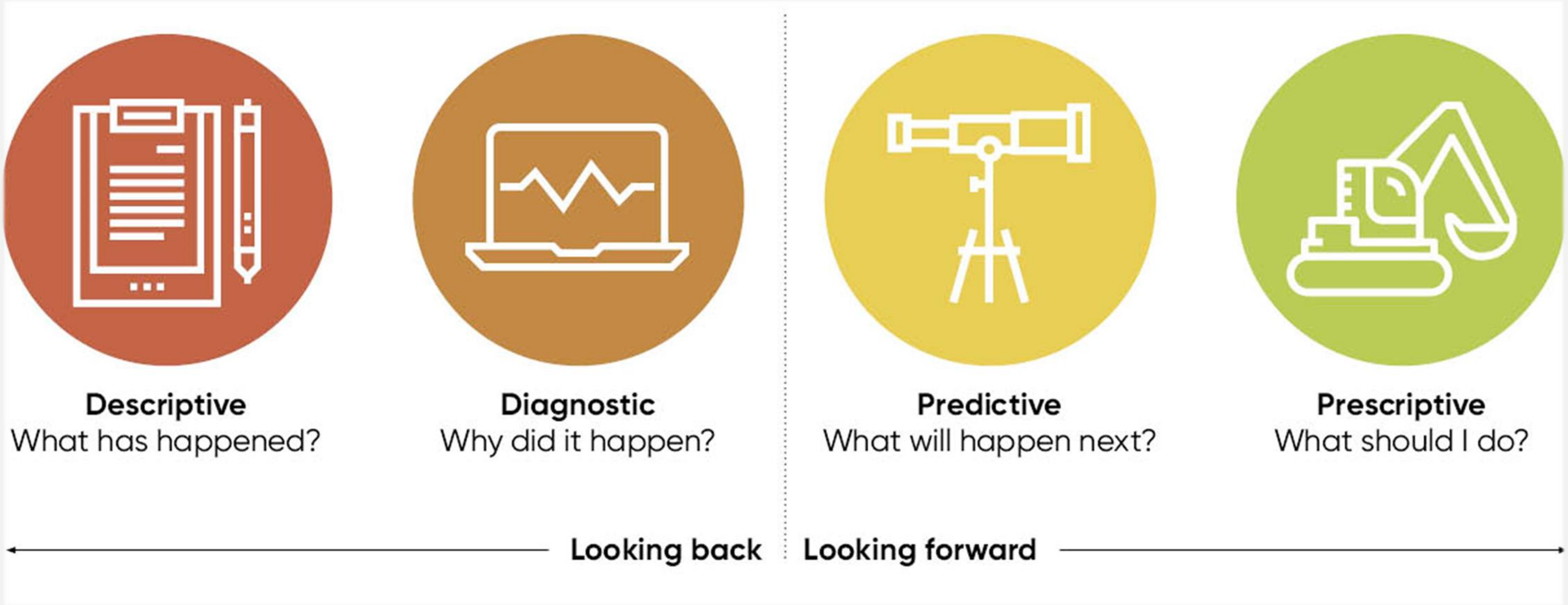
Descriptive
Explains what happened.

Diagnostic
Explains why it happened.

Predictive
Forecasts what might happen.

Prescriptive
Recommends an action based on the forecast.

Business Analytics



Descriptive
What has happened?

Diagnostic
Why did it happen?

Predictive
What will happen next?

Prescriptive
What should I do?

Looking back | **Looking forward**

Hands-on

We will do some Data Analysis using BI Tools (Tableau)

Download Tableau Public

https://public.tableau.com/en-us/s/download

Data Analysis

- ❑ Data in Raw format might not help.
- ❑ Data transformation through calculation help to do :
  - ❑ Draw better insights
  - ❑ Generate Report
  - ❑ Data Driven Decision
  - ❑ Self Serving Data

Get Dataset (Super Store Sales Data)

https://drive.google.com/file/d/131VI-hVyeLFRwkFNqaa01N_Rk8NHlZvs/view

Business Questions

❑ What is the growth of various Sub-Categories over 4 years?

❑ Which category in each segment is yielding more profit?

❑ Monthly fluctuations in sales in various years?

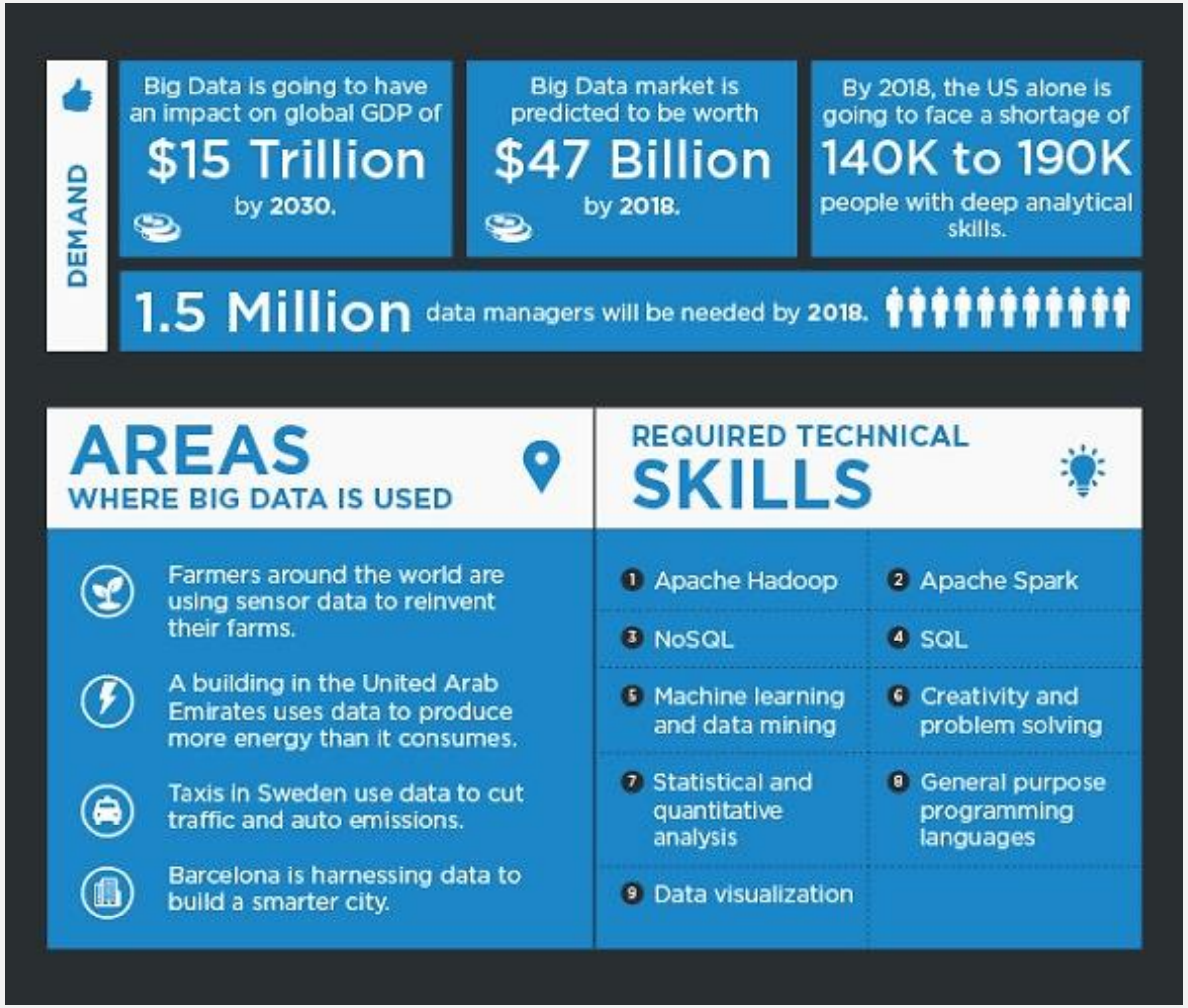
❑ Running total of sales in each year?

❑ Rank sub-categories based on Quantity sold and compare their profits.

❑ Find the average discount to Sales ratio for each sub-category in different regions.

❑ What is the average order to ship time for various sub-categories?

Big Data – Career Path



DEMAND

Big Data is going to have an impact on global GDP of **$15 Trillion** by 2030.

Big Data market is predicted to be worth **$47 Billion** by 2018.

By 2018, the US alone is going to face a shortage of **140K to 190K** people with deep analytical skills.

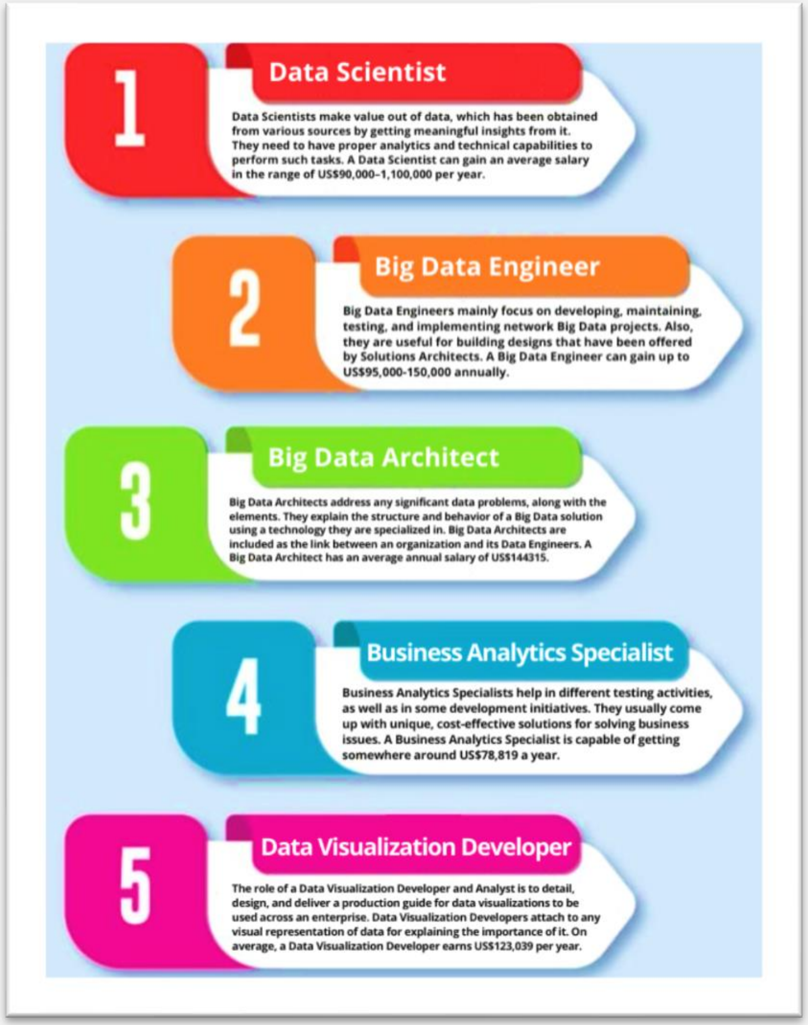
1.5 Million data managers will be needed by 2018.

AREAS
WHERE BIG DATA IS USED

- Farmers around the world are using sensor data to reinvent their farms.
- A building in the United Arab Emirates uses data to produce more energy than it consumes.
- Taxis in Sweden use data to cut traffic and auto emissions.
- Barcelona is harnessing data to build a smarter city.

REQUIRED TECHNICAL SKILLS

1. Apache Hadoop
2. Apache Spark
3. NoSQL
4. SQL
5. Machine learning and data mining
6. Creativity and problem solving
7. Statistical and quantitative analysis
8. General purpose programming languages
9. Data visualization



1 **Data Scientist**

Data Scientists make value out of data, which has been obtained from various sources by getting meaningful insights from it. They need to have proper analytics and technical capabilities to perform such tasks. A Data Scientist can gain an average salary in the range of US$90,000–1,100,000 per year.

2 **Big Data Engineer**

Big Data Engineers mainly focus on developing, maintaining, testing, and implementing network Big Data projects. Also, they are useful for building designs that have been offered by Solutions Architects. A Big Data Engineer can gain up to US$95,000-150,000 annually.

3 **Big Data Architect**

Big Data Architects address any significant data problems, along with the elements. They explain the structure and behavior of a Big Data solution using a technology they are specialized in. Big Data Architects are included as the link between an organization and its Data Engineers. A Big Data Architect has an average annual salary of US$144315.

4 **Business Analytics Specialist**

Business Analytics Specialists help in different testing activities, as well as in some development initiatives. They usually come up with unique, cost-effective solutions for solving business issues. A Business Analytics Specialist is capable of getting somewhere around US$78,819 a year.

5 **Data Visualization Developer**

The role of a Data Visualization Developer and Analyst is to detail, design, and deliver a production guide for data visualizations to be used across an enterprise. Data Visualization Developers attach to any visual representation of data for explaining the importance of it. On average, a Data Visualization Developer earns US$123,039 per year.