Big Data Characteristics, Type of Analytics and Big Data Architecture

Dr. V. K. Patle
Associate Professor
S. o. S. Computer Science & IT
Pt. Ravishankar Shukla University, Raipur (C.G.)

Why is “big data” a “big deal”?

- **Government**
  - Obama administration announced “big data” initiative
  - Many different big data programs launched
- **Private Sector**
  - Walmart handles more than 1 million customer transactions every hour, which is imported into databases estimated to contain more than 2.5 petabytes of data
  - Facebook handles 40 billion photos from its user base.
  - Falcon Credit Card Fraud Detection System protects 2.1 billion active accounts world-wide
- **Science**
  - Large Synoptic Survey Telescope will generate 140 Terabyte of data every 5 days.
  - Biomedical computation like decoding human Genome & personalized medicine
  - Social science revolution
  - -…

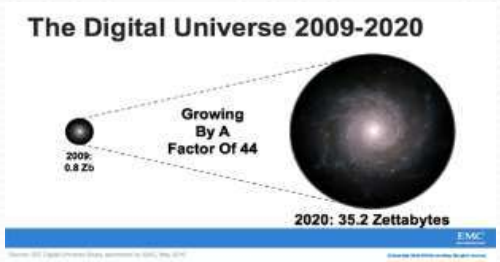
Characteristics of Big Data

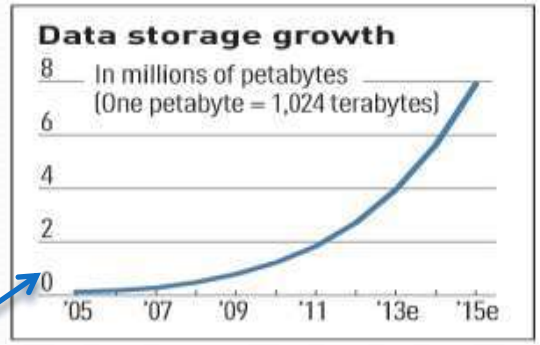
Volume:

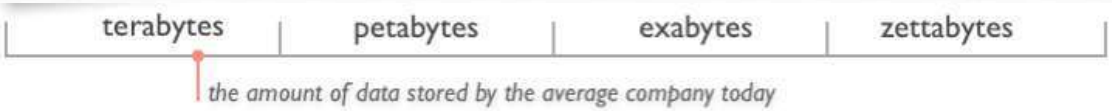
- The quantity of generated and stored data.
- The size of the data determines the value and potential insight, and whether it can be considered big data or not.
- Size of data plays a very crucial role in determining value out of data. Hence, 'Volume' is needed to be considered while dealing with 'Big Data'.
- Organizations collect data from a variety of sources, including business transactions, social media and information from a sensor or machine-to-machine data.
- In the past, storing it would’ve been a problem – but new technologies (such as Hadoop) have eased the burden.

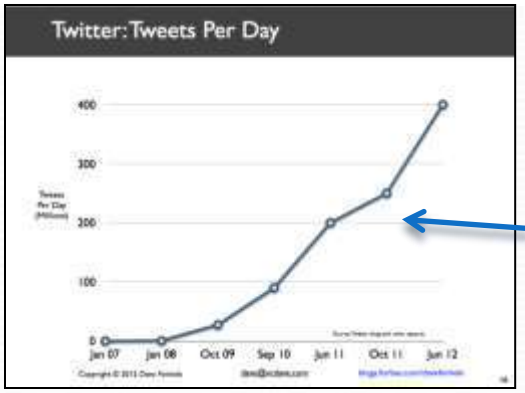
Volume (Scale)

- **Data Volume**
  - 44x increase from 2009 2020
  - From 0.8 zettabytes to 35zb
- Data volume is increasing exponentially









Exponential increase in collected/generated data

Velocity (Speed)

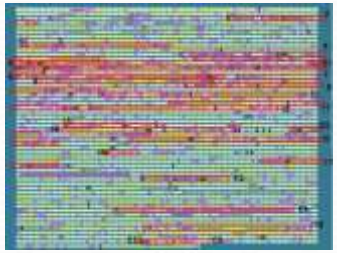
- Data is begin generated fast and need to be processed fast
- Online Data Analytics
- Late decisions ➔ missing opportunities
- **Examples**
  - **E-Promotions:** Based on your current location, your purchase history, what you like ➔ send promotions right now for store next to you
  - **Healthcare monitoring:** sensors monitoring your activities and body ➔ any abnormal measurements require immediate reaction

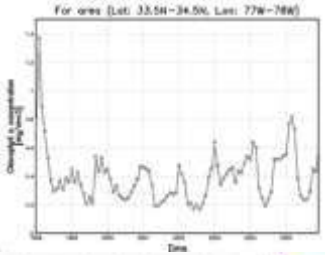
Variety:

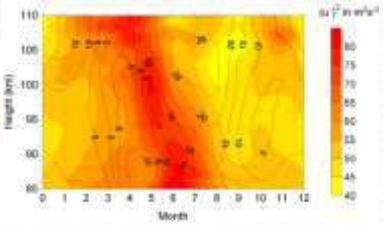
- Variety refers to heterogeneous sources and the type and nature of data, both structured and unstructured.
- Big data draws from text, images, audio, video; plus it completes missing pieces through data fusion.
- Big data systems usually accept and store data closer to its raw state.
- Ideally, any transformations or changes to the raw data will happen in memory at the time of processing.
- Big data seeks to handle potentially useful data regardless of where it's coming from by consolidating all information into a single system.

Variety (Complexity)

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- Graph Data
  - Social Network, Semantic Web (RDF), …
- Streaming Data
  - You can only scan the data once
- A single application can be generating/collecting many types of data
- Big Public Data (online, weather, finance, etc)

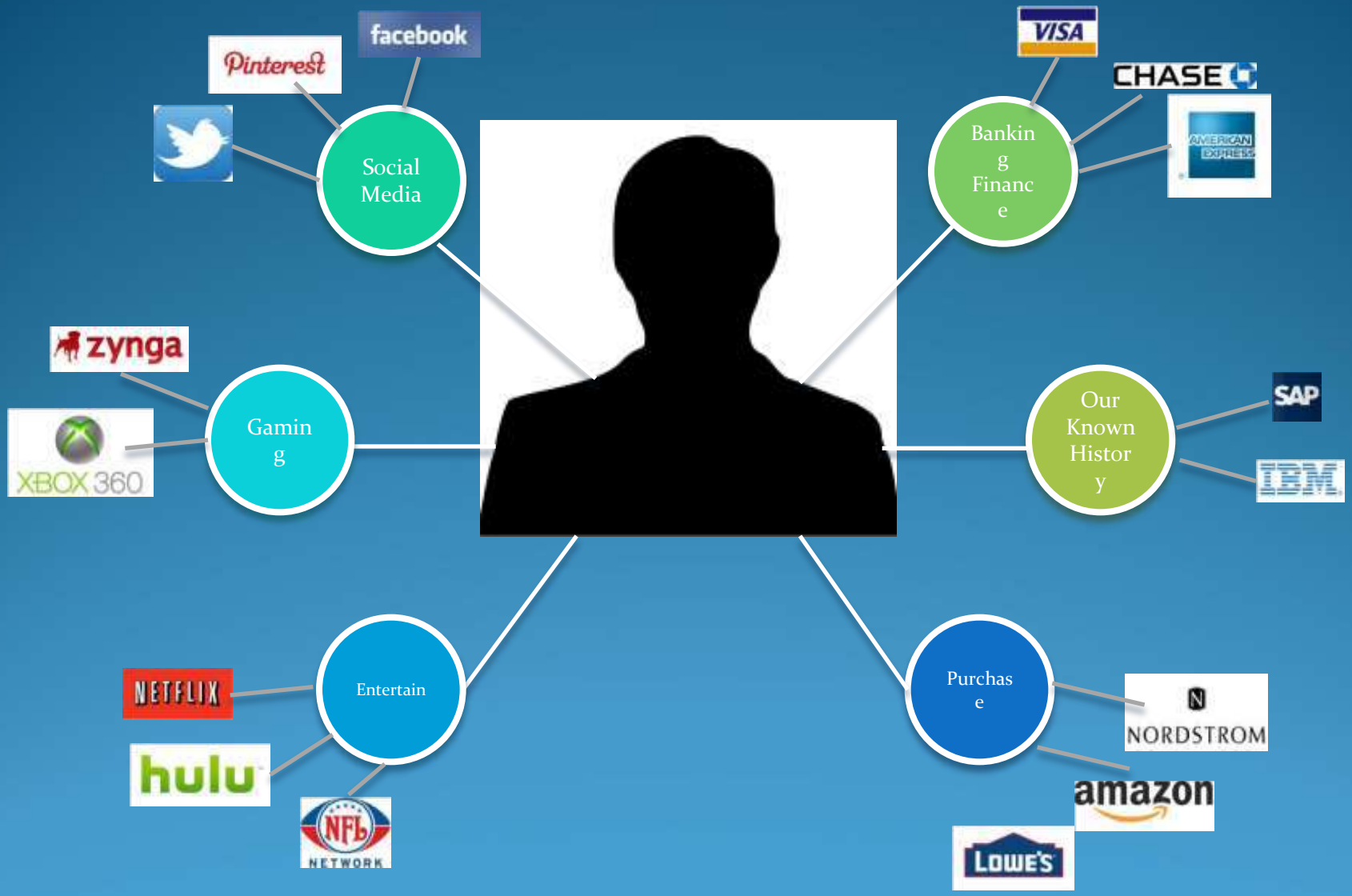




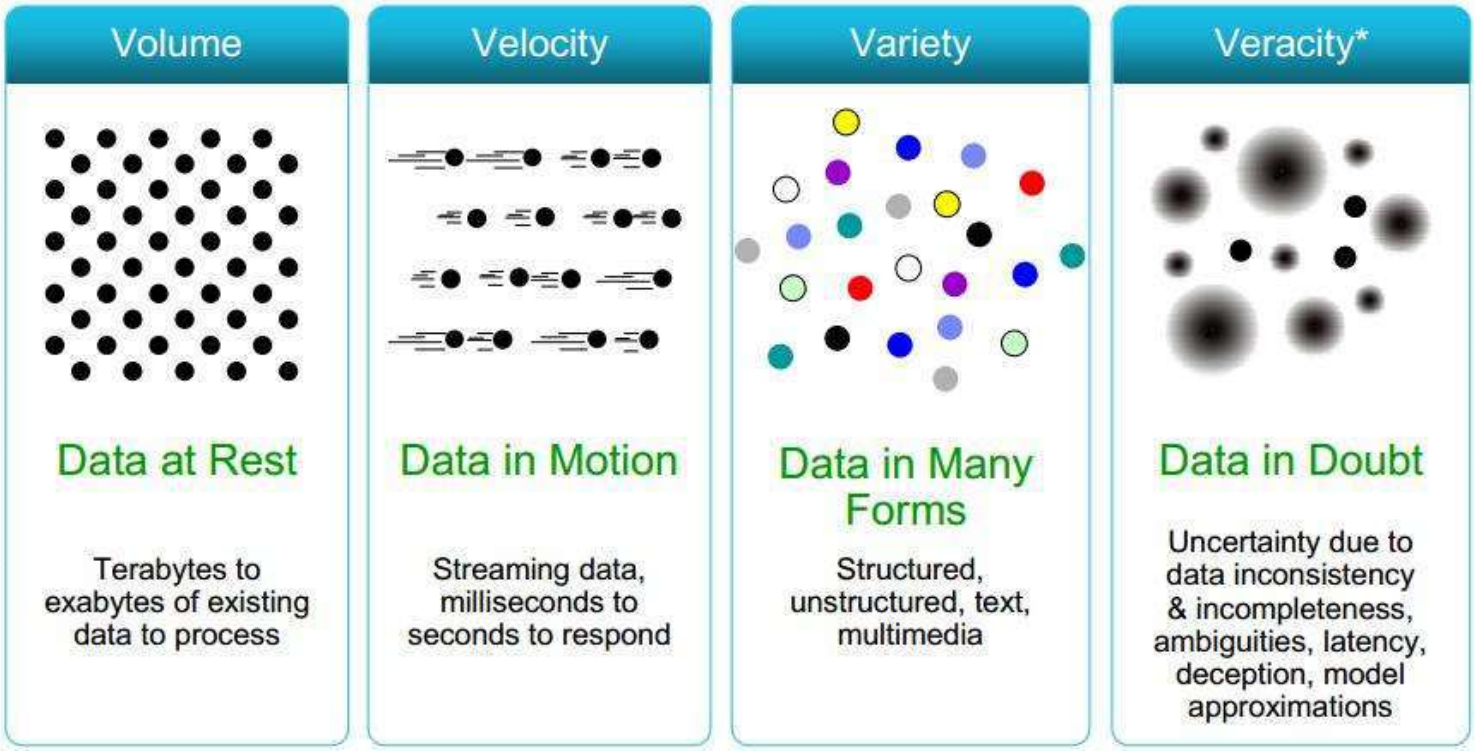


To extract knowledge➔ all these types of data need to linked together

A Single View to the Customer



Some Make it 4V's



Veracity: The variety of sources and the complexity of the processing can lead to challenges in evaluating the quality of the data . Data must be processed with advanced tools (analytics and algorithms) to reveal meaningful information.

Variability: Variation in the data leads to wide variation in quality. Additional resources may be needed to identify, process, or filter low quality data to make it more useful.

Validity: Validity refers to how accurate and correct the data is for its intended use. The benefit from big data analytics is only as good as its underlying data, so you need to adopt good data governance practices to ensure consistent data quality, common definitions, and metadata.

Vulnerability: The rapid development in software applications and failure on the part of system developers to properly analyze program codes before been released to the market increases the chance for data breaches. Data Mining and its related algorithms are an active area which can successfully be applied in analyzing software vulnerability.

Volatility: Due to the velocity and volume of big data, however, its volatility needs to be carefully considered. You now need to establish rules for data currency and availability as well as ensure rapid retrieval of information when required. The costs and complexity of a storage and retrieval process are magnified with big data.

Visualization: big data visualization tools face technical challenges due to limitations of in-memory technology and poor scalability, functionality, and response time. you need different ways of representing data such as data clustering or using tree maps, sunbursts, parallel coordinates, circular network diagrams, or cone trees.

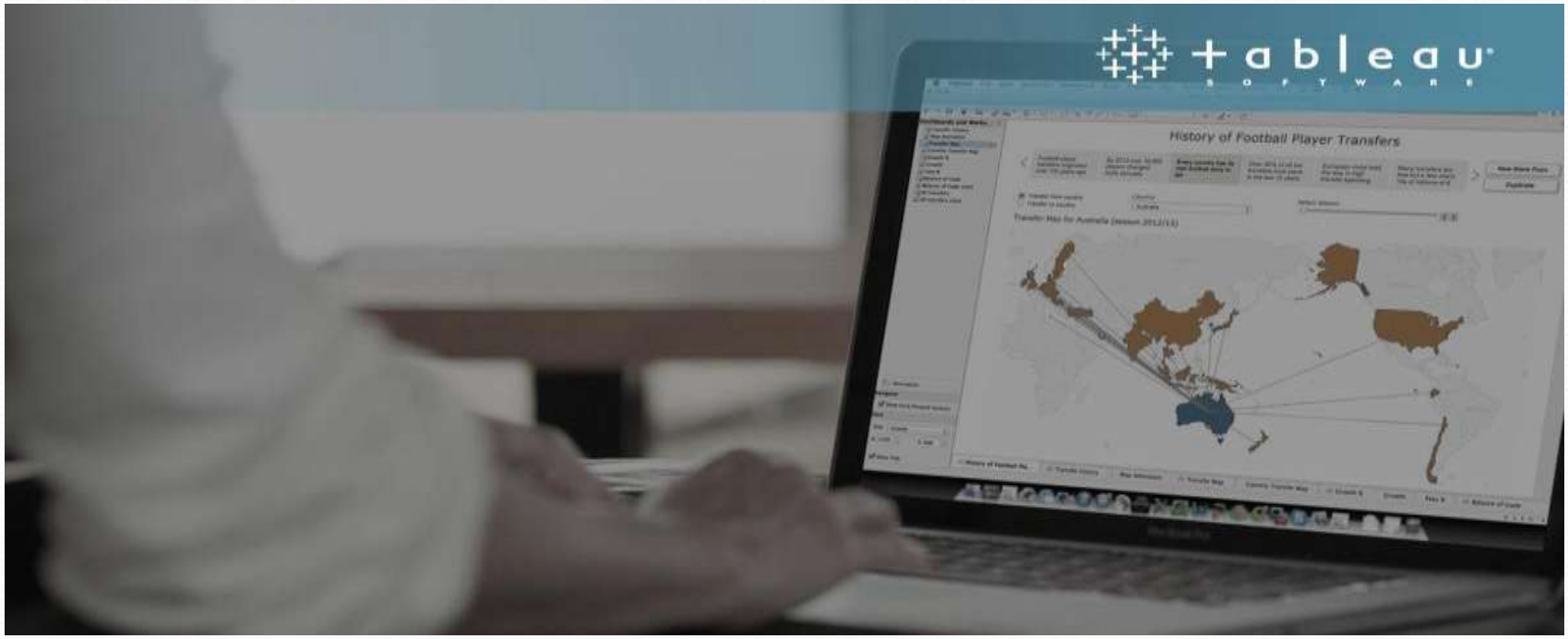
What is Data Visualization?

Data visualization is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly.

Why Tableau for Data Visualization?

Tableau is a powerful, flexible Data Visualization tool that is easy to learn, easy to use, and has powerful libraries for data visualization and presentation.

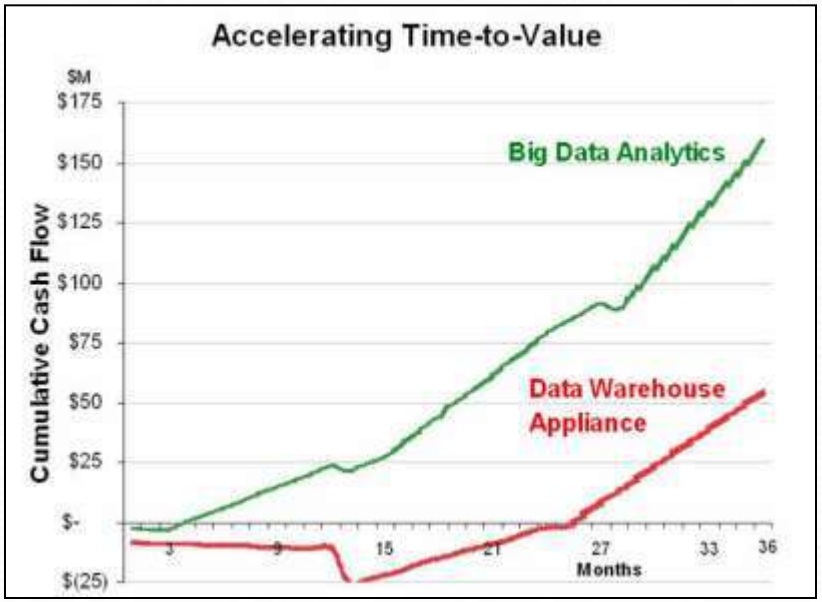
Cost of Ownership	Tableau is a competitively priced software that is available for a trial download.
Versatility	Multi-purpose package that can be used to build an entire application
Big data compatibility	Tableau has become one of the big go-to software programs for Data visualization due to the wide variety of tools it provides and compatibility with Big Data platforms such as Hadoop.



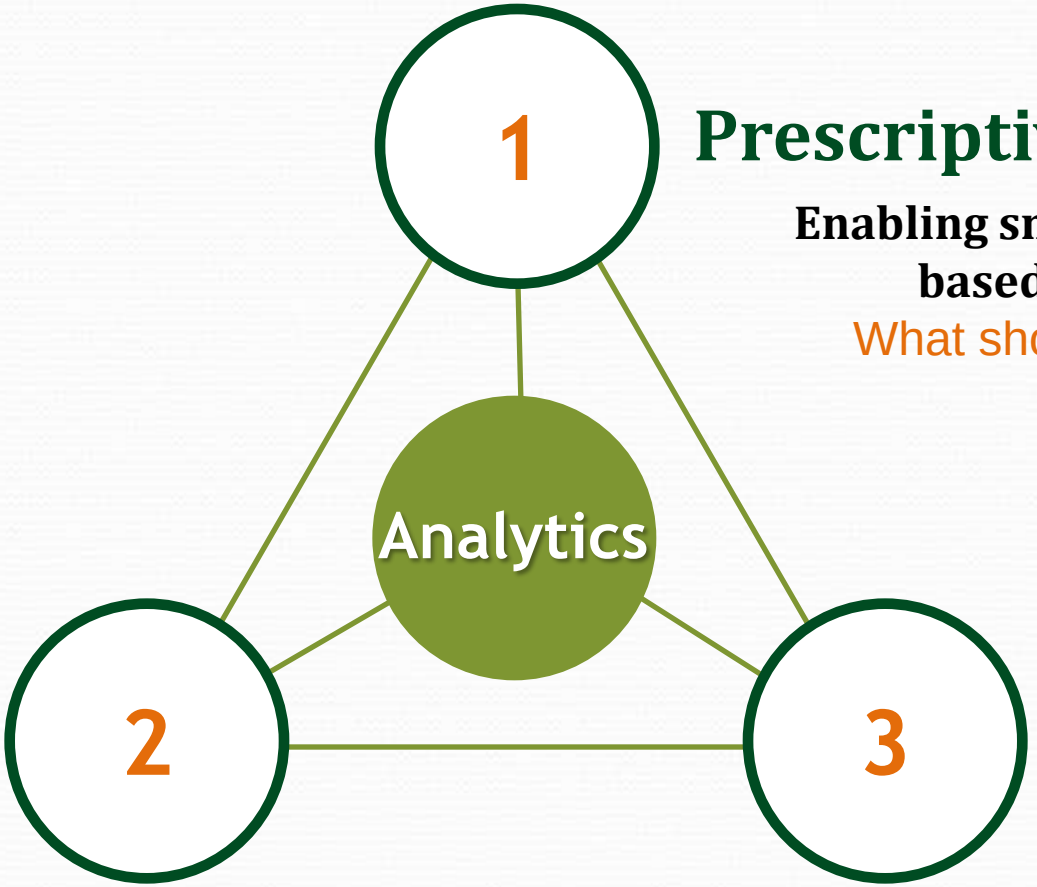
Value: Substantial value can be found in big data, including understanding your customers better, targeting them accordingly, optimizing processes, and improving machine or business performance. You need to understand the potential, along with the more challenging characteristics, before embarking on a big data strategy.

Big Data Analytics

- Big data is more real-time in nature than traditional DW applications
- Traditional DW architectures (e.g. Exadata, Teradata) are not well-suited for big data apps
- Shared nothing, massively parallel processing, scale out architectures are well-suited for big data apps



Types of Analytics



Prescriptive Analytics

Enabling smart decisions based on data

What should we do?

Predictive analytics

Predicting the future based on historical patterns

What could happen?

Descriptive analytics

Mining data to provide business insights

What has happened?

Types of Analytics

Why do airline prices change every hour?

Prescriptive Analytics
advice on possible outcomes

How do grocery cashiers know to hand you coupons you might actually use?

Predictive Analytics
understanding the future

How does Netflix frequently recommend just the right movie?

Descriptive Analytics
insight into the past

Growing Need for Analytics

Generation of Large Amount of Data from Business Transactions

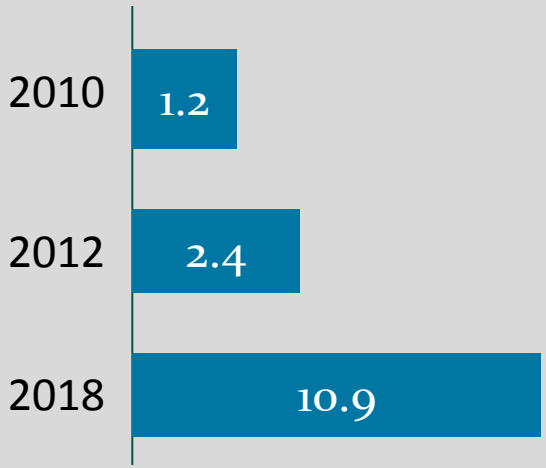
DATA HARNESSING
Companies store each piece of information generated during the business operations and customer interactions.

Data is generated.

Data is analyzed.

Learning from the data is used in the decision making and process optimization.

DATA VOLUMES



Volumes in Trillion GB

Walmart

DID YOU KNOW ?

4 Billion — Number of transactions every year

900 — Number of Stores

10000 -1 lakh — Number of SKUs

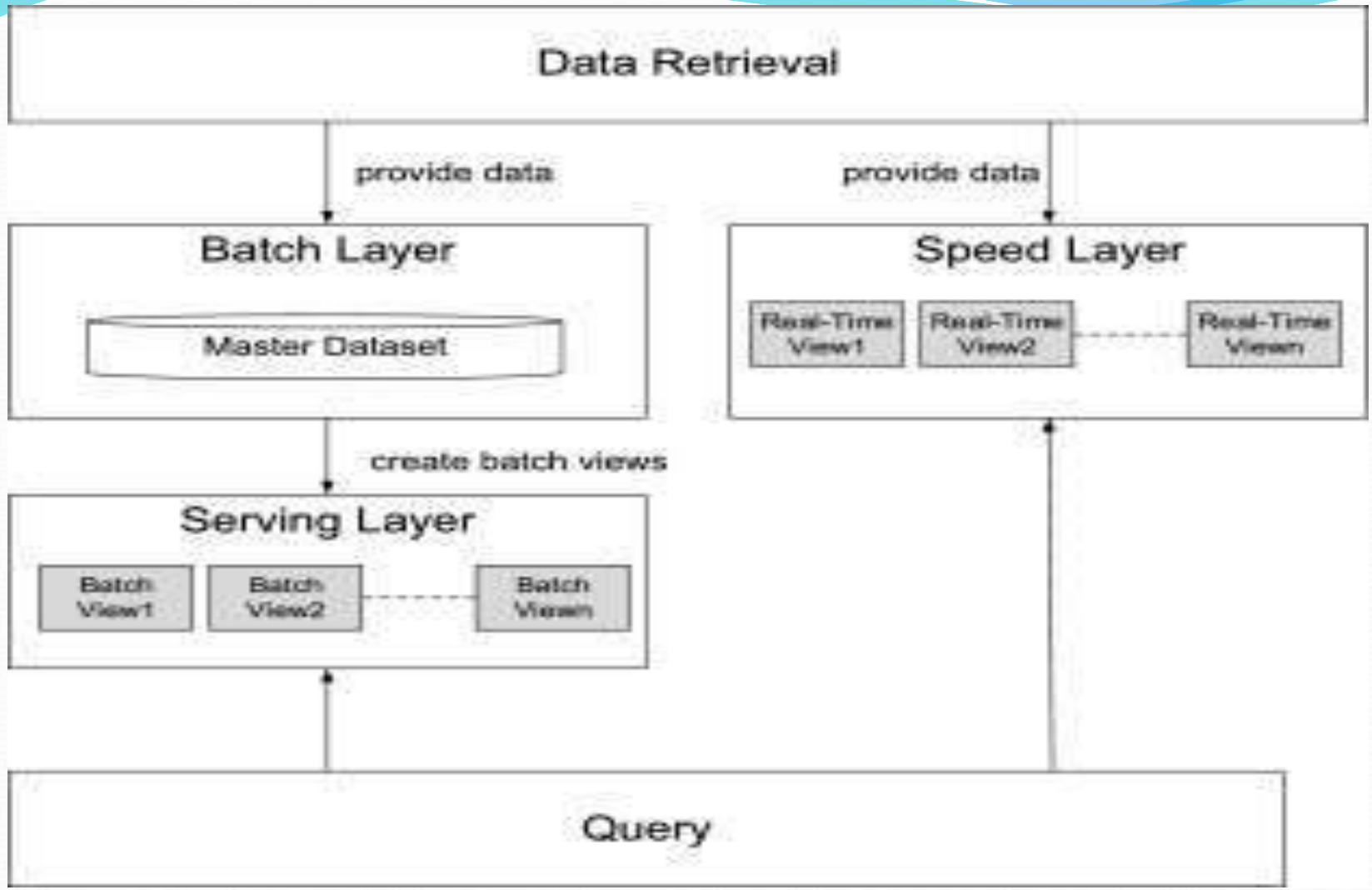
Why Big Data Analytics?

Why is Big Data Analytics important?

Big data analytics helps organizations harness their data and use it to identify new opportunities. That, in turn, leads to smarter business moves, more efficient operations, higher profits and happier customers.

Big Data Architecture

- An appropriate big data architecture design will play a fundamental role to meet the big data processing needs. Several reference architectures are now being proposed to support the design of big data systems.
- The Lambda architecture as defined by Marz [10]. The Lambda architecture is a big data architecture that is designed to satisfy the needs for a robust system that is fault-tolerant, both against hardware failures and human mistakes. Hereby it takes advantage of both batch- and stream-processing methods. In essence, the architecture consists of three layers including batch processing layer, speed (or real-time) processing layer, and serving layer.



https://www.sciencedirect.com/topics/computer-science/big-data-architecture

Conclusion

In this presentation cover the characteristics of big data, Type of V, Type of analytics and Big data Architecture concept. Also cover the Visualization

Thank You