Introduction

- Big Data may well be the Next Big Thing in the IT world.
- Big data burst upon the scene in the first decade of the 21st century.
- The first organizations to embrace it were online and startup firms. Firms like Google, eBay, LinkedIn, and Facebook were built around big data from the beginning.
- Like many new information technologies, big data can bring about dramatic cost reductions, substantial improvements in the time required to perform a computing task, or new product and service offerings.

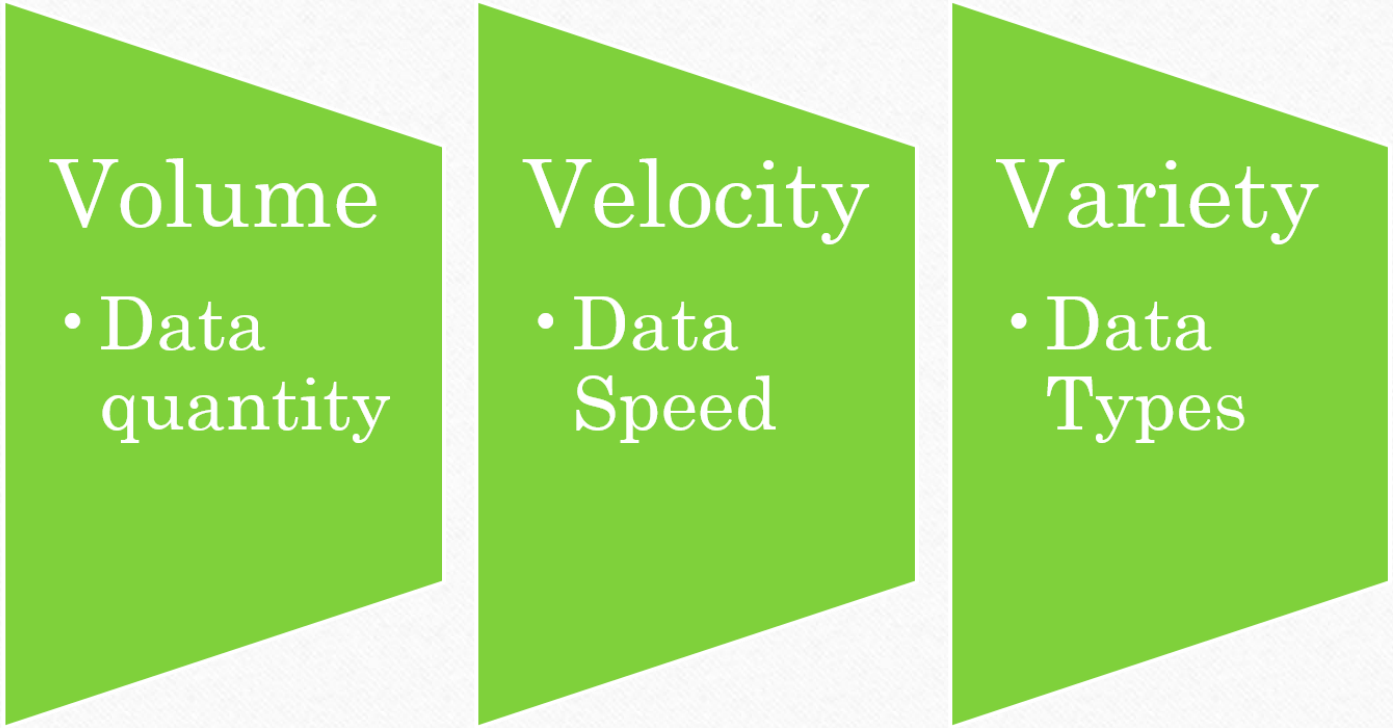
What is BIG DATA?

- 'Big Data' is similar to 'small data', but bigger in size but having data bigger it requires different approaches:
  - Techniques, tools and architecture
- an aim to solve new problems or old problems in a better way
- Big Data generates value from the storage and processing of very large quantities of digital information that cannot be analyzed with traditional computing techniques.

What is BIG DATA

- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion photos from its user base.
- Decoding the human genome originally took 10years to process; now it can be achieved in one week.

Three Characteristics of Big Data V3s



1st Character of Big Data
Volume

- A typical PC might have had 10 gigabytes of storage in 2000.
- Today, Facebook ingests 500 terabytes of new data every day.
- Boeing 737 will generate 240 terabytes of flight data during a single flight across the US.
- The smart phones, the data they create and consume; sensors embedded into everyday objects will soon result in billions of new, constantly-updated data feeds containing environmental, location, and other information, including video.

2nd Character of Big Data

Velocity

- Clickstreams and ad impressions capture user behavior at millions of events per second
- high-frequency stock trading algorithms reflect market changes within microseconds
- machine to machine processes exchange data between billions of devices
- infrastructure and sensors generate massive log data in real-time
- on-line gaming systems support millions of concurrent users, each producing multiple inputs per second.

3rd Character of Big Data
Variety

- Big Data isn't just numbers, dates, and strings. Big Data is also geospatial data, 3D data, audio and video, and unstructured text, including log files and social media.
- Traditional database systems were designed to address smaller volumes of structured data, fewer updates or a predictable, consistent data structure.
- Big Data analysis includes different types of data

Storing Big Data

❖ **Analyzing your data characteristics**

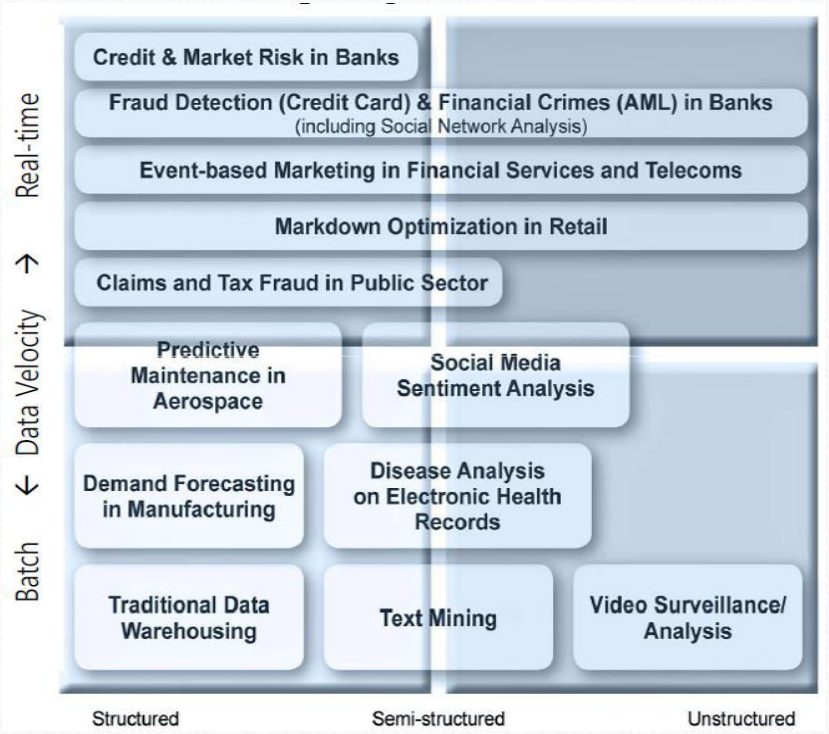
- Selecting data sources for analysis
- Eliminating redundant data
- Establishing the role of NoSQL

❖ **Overview of Big Data stores**

- Data models: key value, graph, document, column-family
- Hadoop Distributed File System

The Structure of Big Data

- Structured
  - Most traditional data sources
- Semi-structured
  - Many sources of big data
- Unstructured
  - Video data, audio data

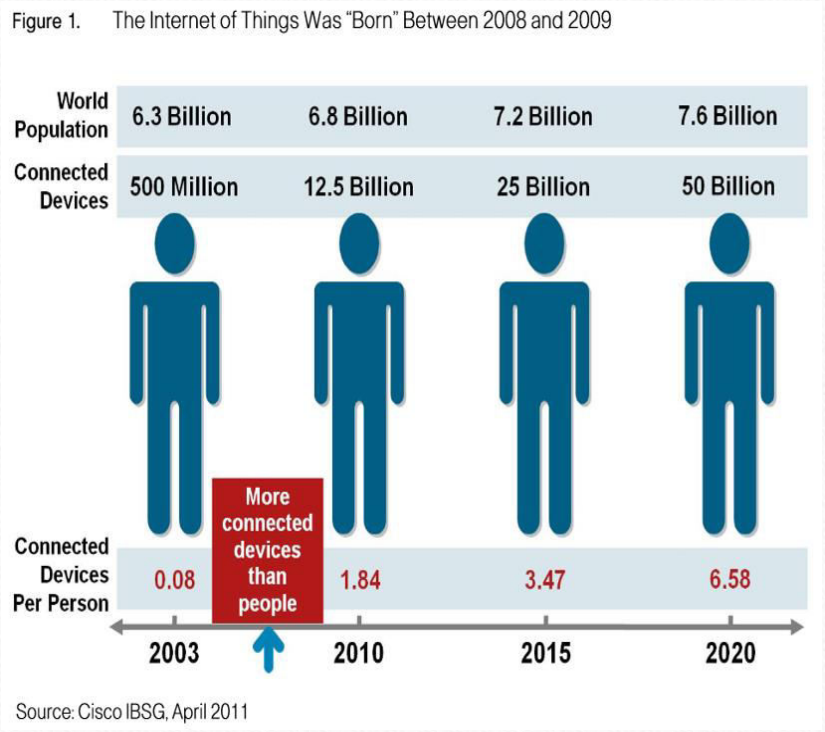


Why Big Data

- Growth of Big Data is needed
  - Increase of storage capacities
  - Increase of processing power
  - Availability of data(different data types)
  - Every day we create 2.5 quintillion bytes of data; 90% of the data in the world today has been created in the last two years alone

Why Big Data

- FB generates 10TB daily
- Twitter generates 7TB of dataDaily
- IBM claims 90% of today's stored data was generated in just the last two years.



Applications

- Smarter Healthcare
- Traffic Control
- Manufacturing
- Multi-channel sales
- Telecom
- Trading Analytics
- Search Quality

NoSQL

- NoSQL is an approach to database design that can accomodate a wide variety of data models, including key-value, document, columnar and graph formats.
- NoSQL, which stand for "not only SQL," is an alternative to traditional relational databases in which data is placed in tables and data schema is carefully designed before the database is built.
- NoSQL databases are especially useful for working with large sets of distributed data.

Major issues of Relational Databases

- Relational databases were designed for tabular data, with a consistent structure and a fixed schema. They work best for problems that are well defined at the outset. But not good to manage unstructued data.
- **A Large Number of JOINs**
- When you utilize queries that join many different tables, there's an explosion of complexity and computing resource consumption. This results in a corresponding increase in query response times.
- **Frequent Schema Changes**
- **Slow-Running Queries (Despite Extensive Tuning)**

Why NoSQL

- Document Oriented Storage: Data is stored in the form of JSON style documents.
- Replication and high availability
- Auto-sharding
- Rich queries
- Fast in-place updates

Characterisics of NoSQL

- It's more than rows in tables—NoSQL systems store and retrieve data from many formats: key-value stores, graph databases, column-family (Bigtable) stores, document stores, and even rows in tables.
- It's free of joins—NoSQL systems allow you to extract your data using simple interfaces without joins.
- It's schema-free—NoSQL systems allow you to drag-and-drop your data into a folder and then query it without creating an entity-relational model.
- It works on many processors—NoSQL systems allow you to store your database on multiple processors and maintain high-speed performance.
- It uses shared-nothing commodity computers—Most (but not all) NoSQL systems leverage low-cost commodity processors that have separate RAM and disk.
- It supports linear scalability—When you add more processors, you get a consistent increase in performance.

Data Models of NoSQL

- Key-value datamodel
- Document data model
- Column Family data model
- Graph data model

- MongoDB is a _____ database that provides high performance, high availability, and easy scalability.
  a) graph
  b) key value
  c) document
  ~~d) all of the mentioned~~

- What does "Velocity" in Big Data mean?

a) Speed of input data generation

b) Speed of individual machine processors

c) Speed of ONLY storing data

d) Speed of storing and processing data

Thank You.