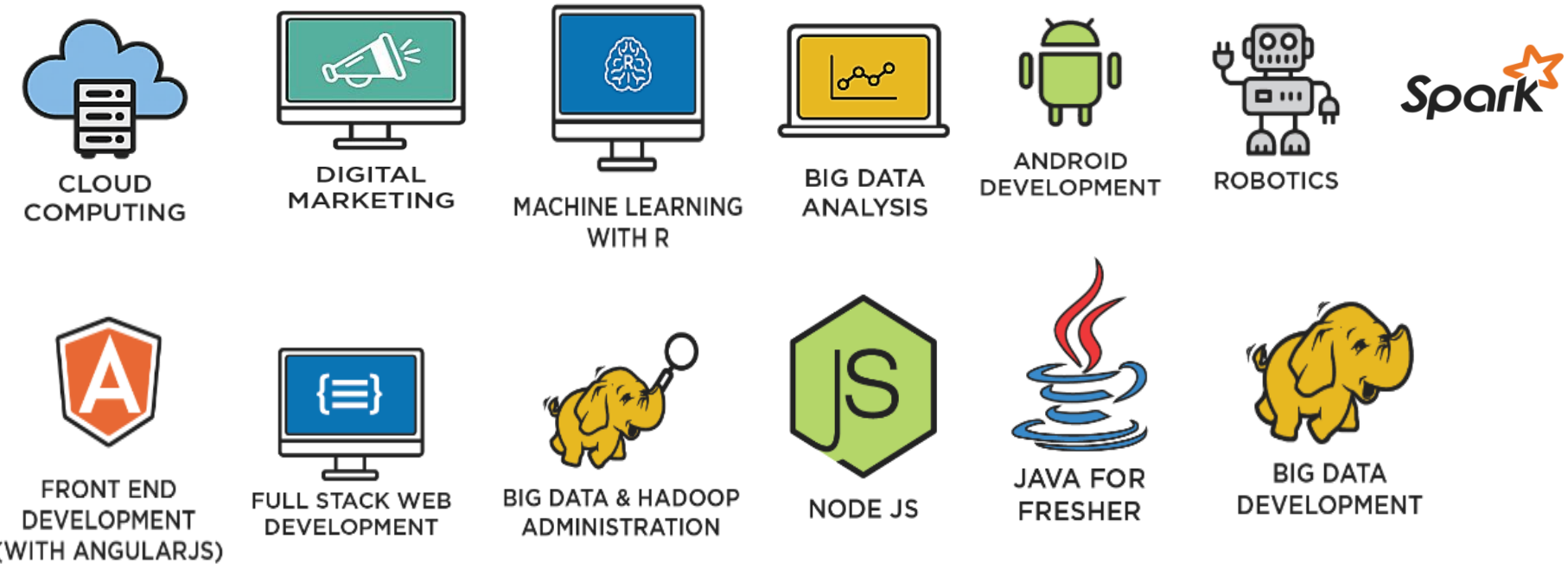

ACADGILD
LEARN. DO. EARN
SPARK
Spark

About AcadGild

ACADGILD

- AcadGild is a technology education start-up which provides online courses in latest technologies.



- AcadGild was founded by Mr. Vinod Dham, Father of Pentium Processors.
- Our aim is to provide millions of high school graduates, college graduates and working professionals, skills to make them ready for jobs.

ACADGILD

Introduction to Big Data

Session 1

LEARN. DO. EARN

Agenda

ACADGILD

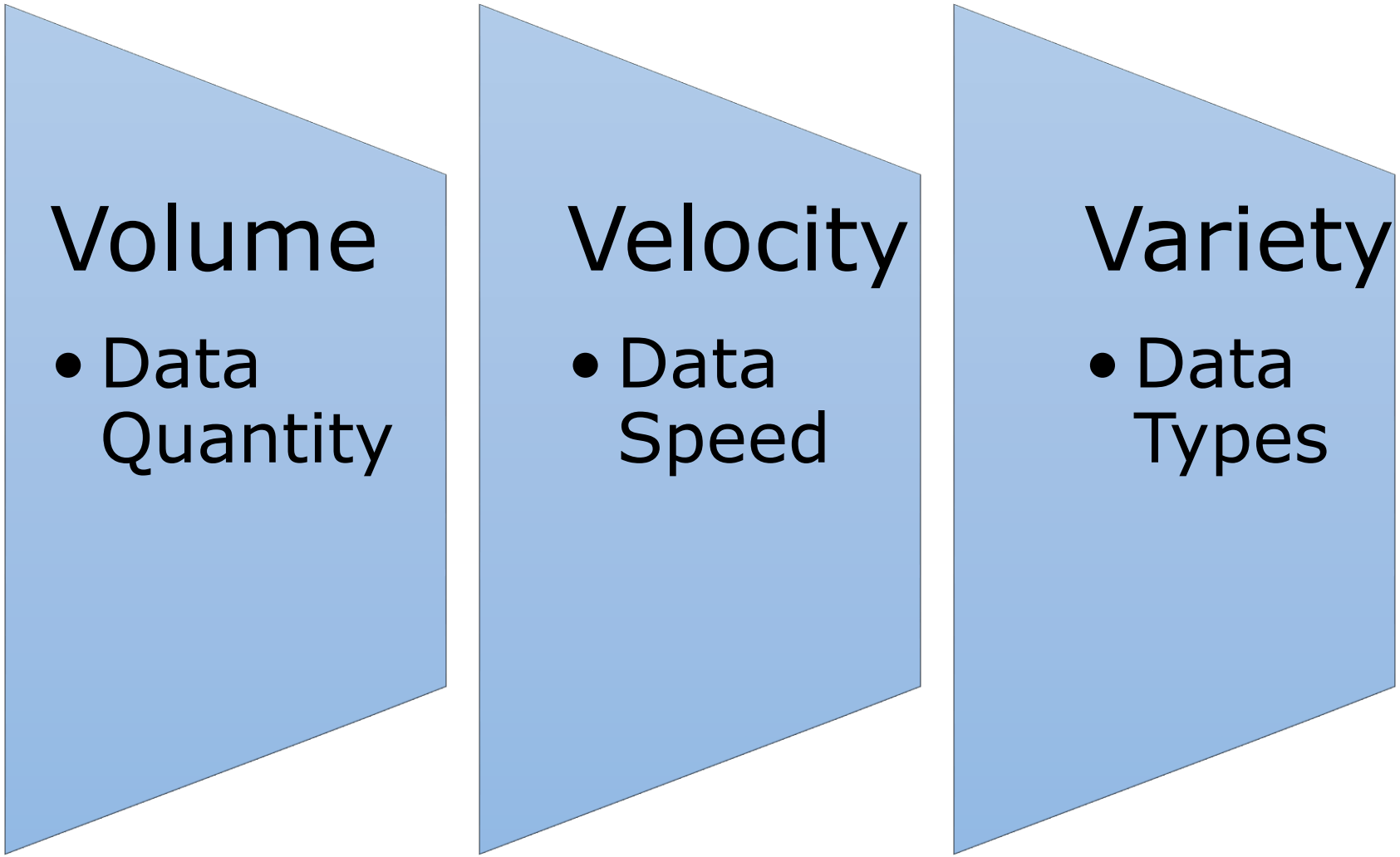
Sl. No.	Agenda Topic
1	Overview of Big Data
2	Characteristics of Big Data
3	Types of Data
4	Sources of Big Data
5	Big Data Examples
6	Scaling
7	Hadoop
8	Hadoop Ecosystem
9	Streaming Data
10	Batch Vs. Streaming Data Processing
11	Streaming Data Processing

Overview of Big Data

- Big data is data that exceeds the processing capacity of conventional database systems.
- Big data is too big in volume, moves too fast in velocity, or doesn’t fit the structures of your database architectures. To gain value from this data, you must choose an alternative way to process it.

Characteristics of Big Data

ACADGILD

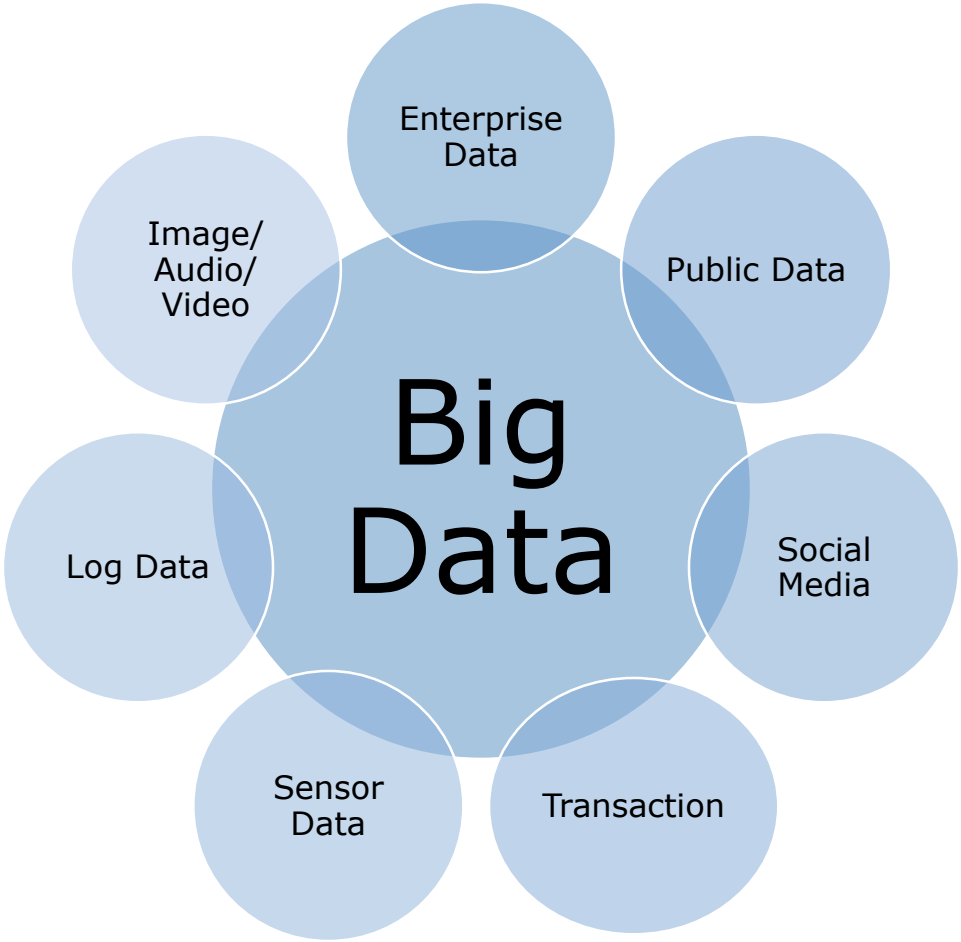


Types of Data

- Structured
  - Most traditional data sources
    - Database, Data warehouse, Enterprise systems
- Semi-structured
  - Many sources of big data
    - XML, Email, EDI
- Unstructured
  - Video data, audio data
    - Analog data, GPS tracking, Audio/Video streams, Text files

Sources of Big Data

ACADGILD



Big Data Examples

- The number of web pages indexed by Google, which were around one million in 1998, have exceeded one trillion in 2008, and its expansion is accelerated by appearance of the social networks.
- Walmart handles more than 1 million customer transactions every hour.
- Facebook handles 40 billion photos from its user base
- More than 65 billion devices were connected to the Internet by 2010, and this number will go up to 230 billion by 2020.

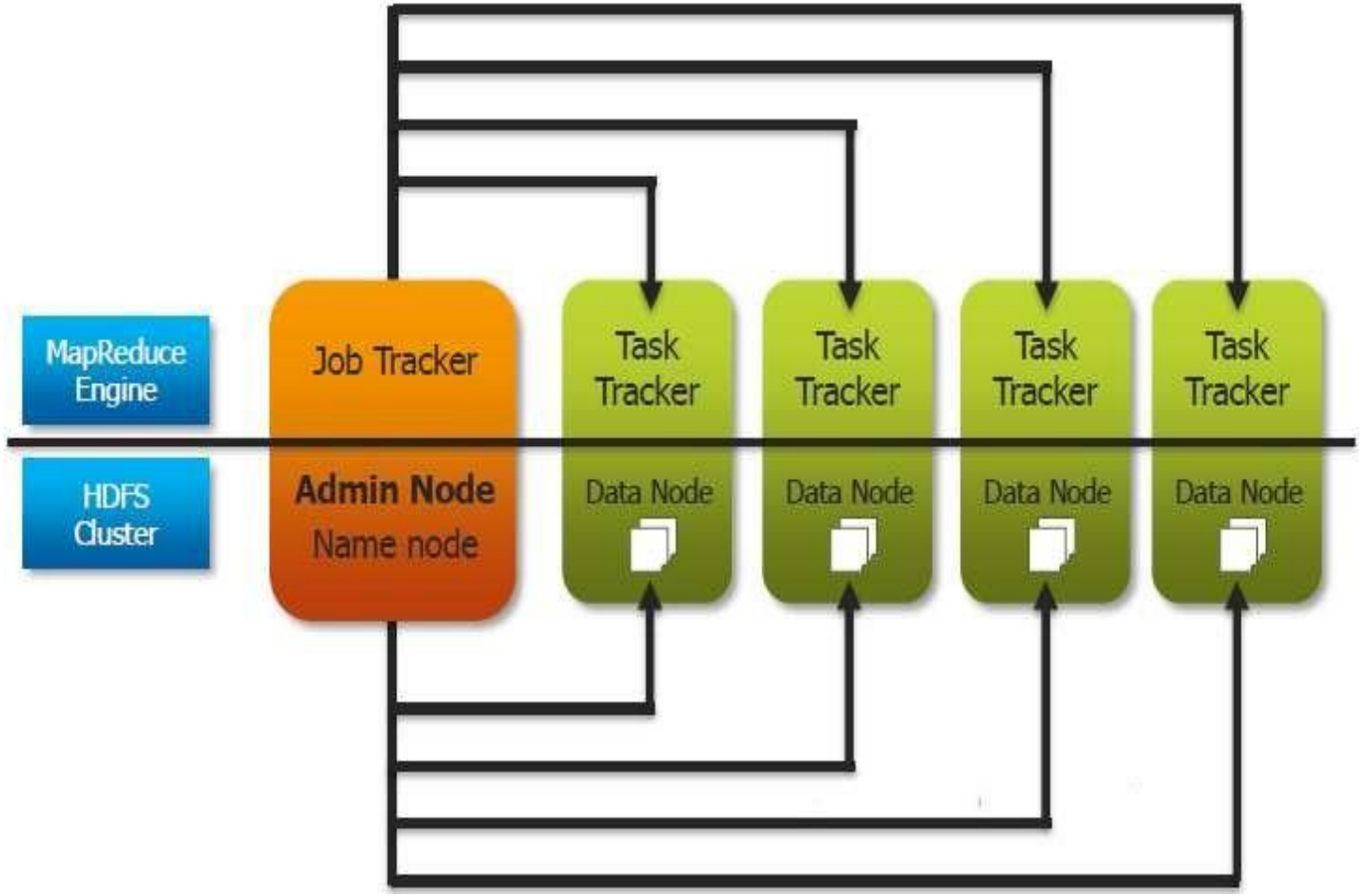
Scaling

- Scaling is one of the issue with conventional systems
  - Scale up or scale vertically: adding resources to a single node in a system
    - more expensive
  - Scale out or scale horizontally: adding more nodes to a system.
    - more challenging for fault tolerance and software development

Hadoop

ACADGILD

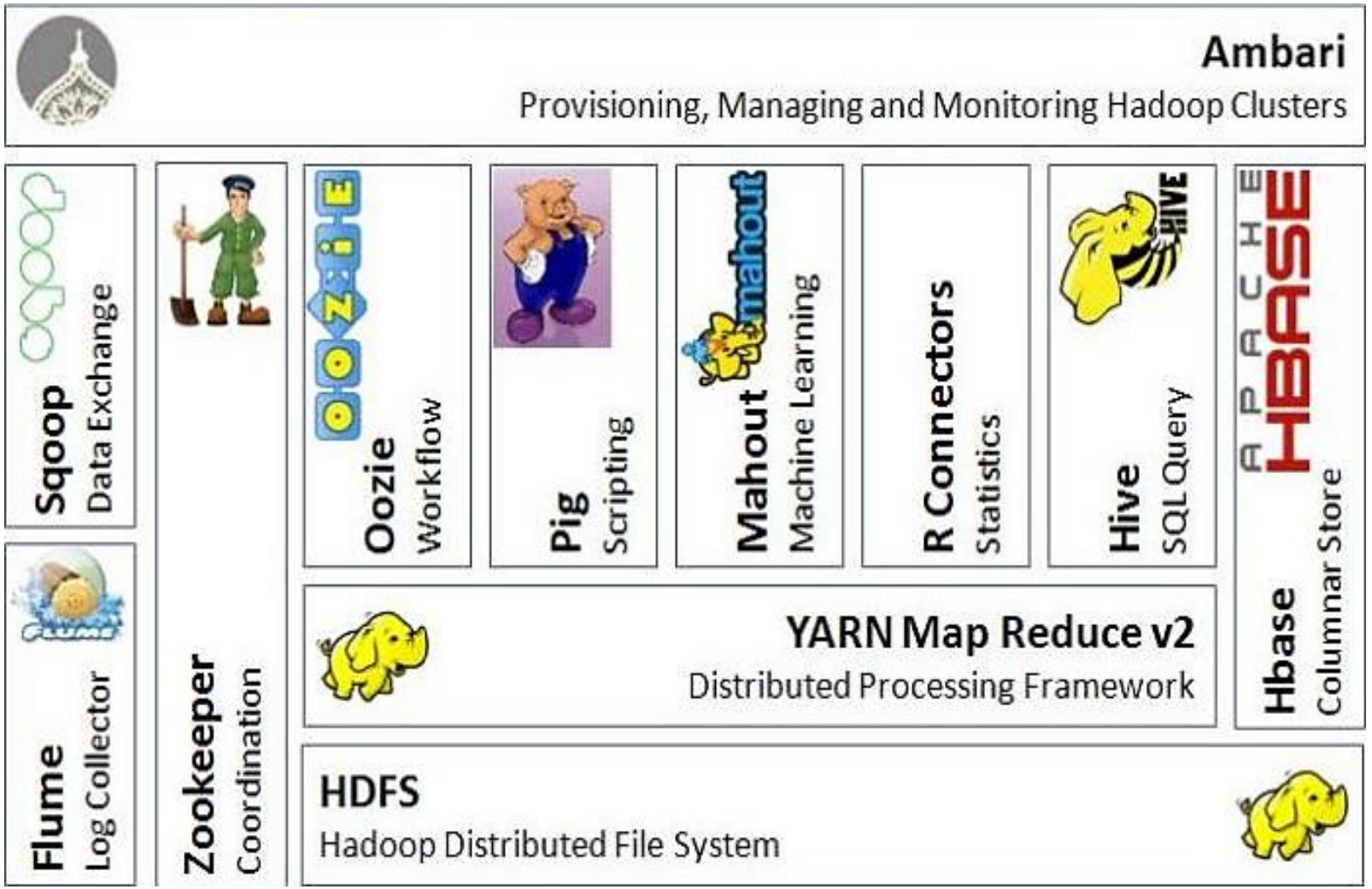
- Apache Hadoop is a framework that allows for the distributed processing of large data sets across clusters of commodity computers using a simple programming model.



Hadoop Ecosystem

ACADGILD

- The main focus of Hadoop ecosystem is on batch processing
- Each requirement needs a specialized tool



Streaming Data

- Data that continuously arrives into the system and needs to be processed before the next interval or its expiry
- Use cases:
  - Complex event processing (CEP) systems
    - event-by-event processing and aggregation
  - Network monitoring
  - Intelligence and surveillance
  - Risk management

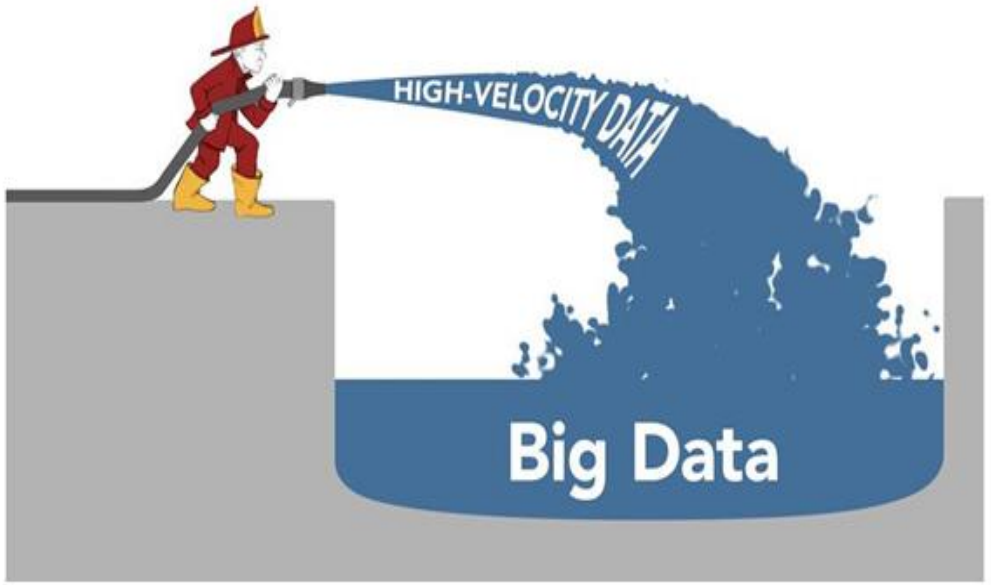
Streaming Data (Contd.)

- E-commerce
- Fraud detection
- Smart order routing
- Transaction cost analysis
- Pricing and analytics
- Market data management
- Algorithmic trading

Streaming Data (Contd.)

ACADGILD

- Data is being created at unprecedented rates
  - Exponential data growth from mobile, web, social
  - Connected devices: 9B in 2012 to 50B by 2020
  - Over 1 trillion sensors by 2020
  - Datacenter IP traffic growing at CAGR of 25%



Batch Vs. Streaming Data Processing

ACADGILD

Streaming Data

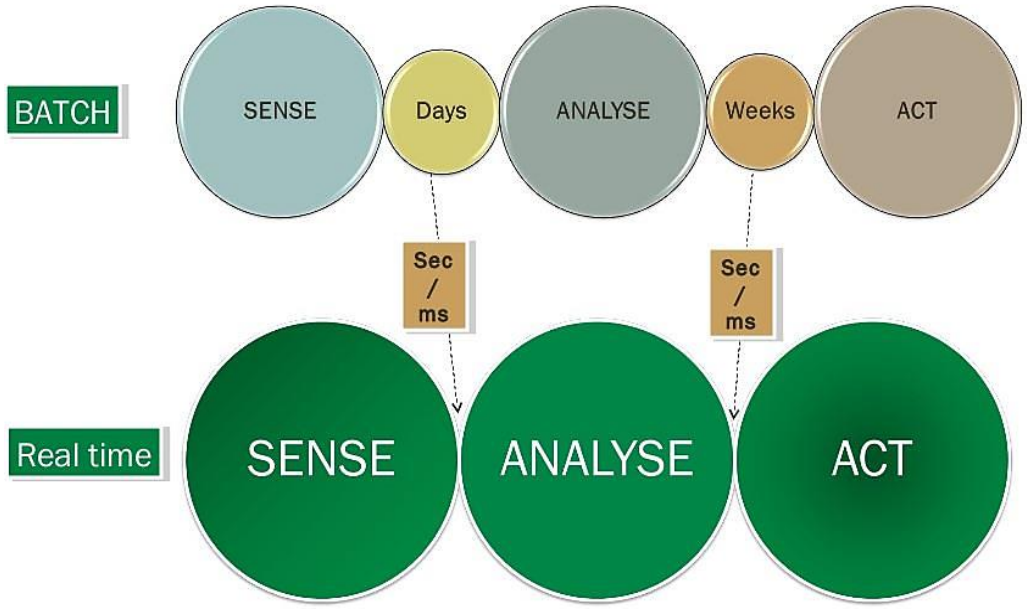
Immediate processing of data after the event occurs

- Real-Time
- Performance and Volume Challenges
- Use Cases: Operations & Analytics

Big Data

Processes after collection and storage of data at a **scheduled time** when a **sufficient amount** of data has been accumulated

- Batch process
- Volume Challenge
- Use Cases: Analytics



Streaming Data Processing

ACADGILD

- Apache Spark and Apache Storm are the Streaming data tools in Hadoop ecosystem

THANK YOU

Email us at - help@acadgild.com