Debre Markos University

Department of Computer Science

Selected Topics in Computer Science

Chapter - Three

Big Data and AI

Compile By

Debalkew G. (MSc.)

Mar-2017

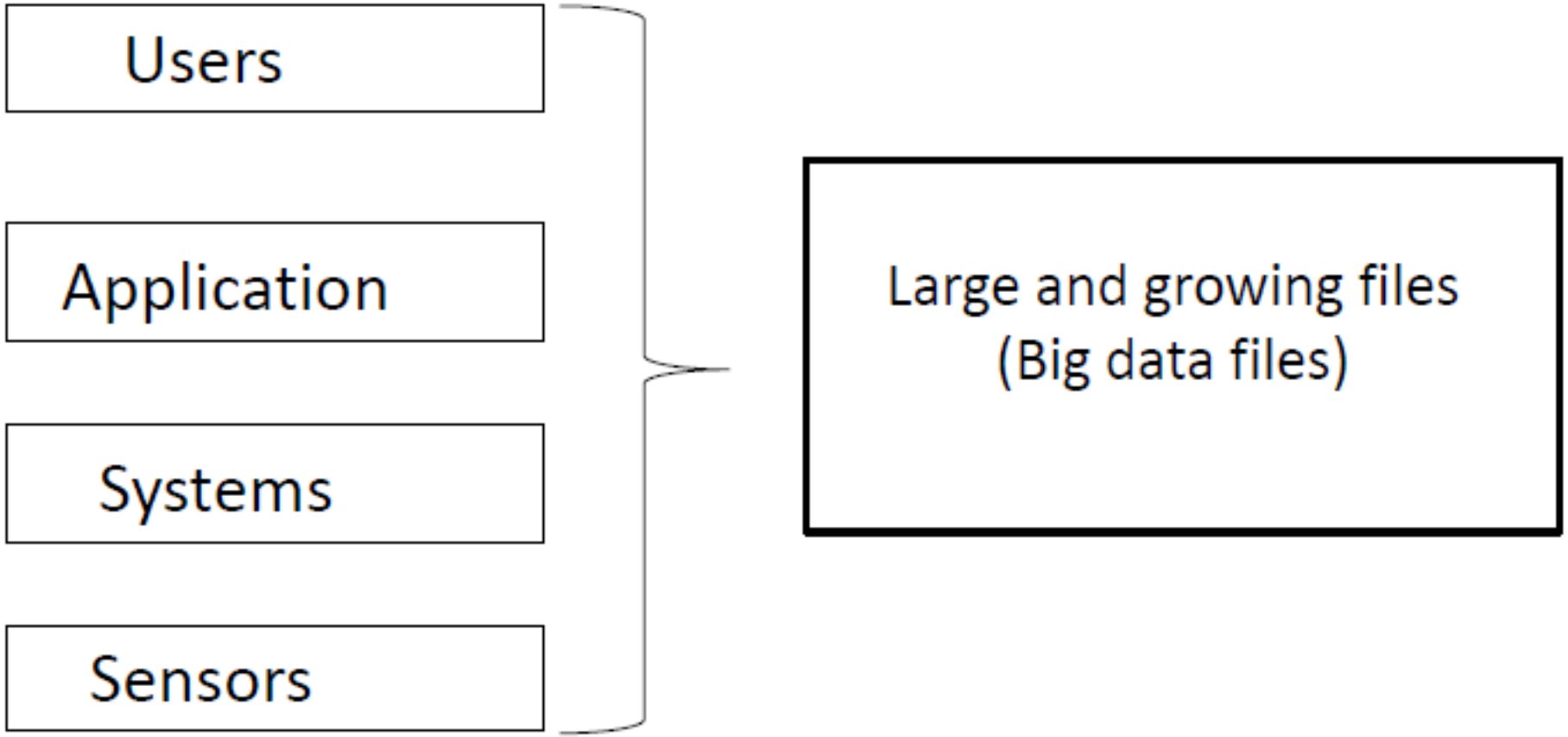
What is Big Data

- big data is larger, more complex data sets, especially from new data sources. These data sets are so voluminous that traditional data processing software just can’t manage them. But these massive volumes of data can be used to address business problems you wouldn’t have been able to tackle before.
- Big data is data that contains greater variety arriving in increasing volumes and with ever higher velocity.

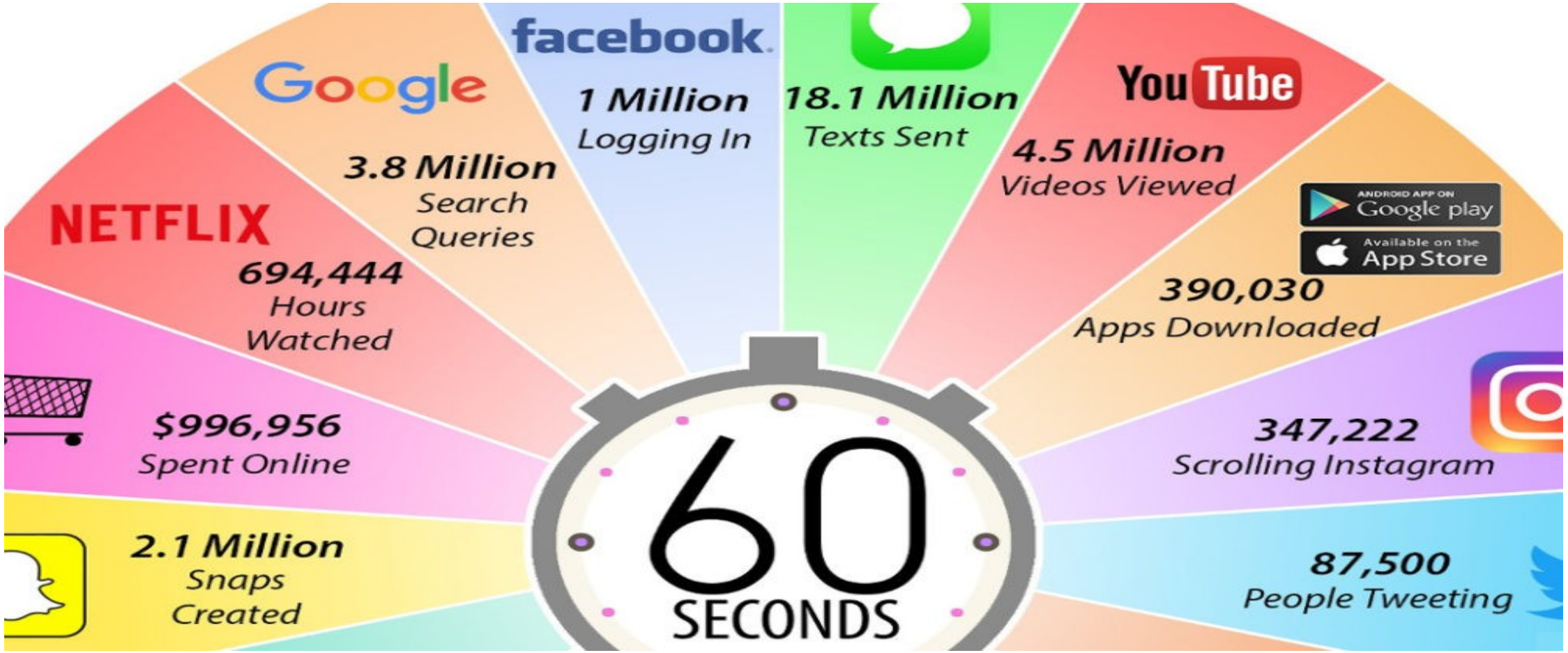
What Comes Under Big Data?

- Big data involves the data produced by different devices and applications. Given below are some of the fields that come under the umbrella of Big Data.
- **Black Box Data** : It is a component of helicopter, airplanes, and jets, etc. It captures voices of the flight crew, recordings of microphones and earphones, and the performance information of the aircraft.
- **Social Media Data** : Social media such as Facebook,Instagram,telgram and Twitter hold information and the views posted by millions of people across the globe.
- **Stock Exchange Data** : The stock exchange data holds information about the 'buy' and 'sell' decisions made on a share of different companies made by the customers.
- **Transport Data** : Transport data includes model, capacity, distance and availability of a vehicle.
- **Search Engine Data** : Search engines retrieve lots of data from different databases.

Big Data sources



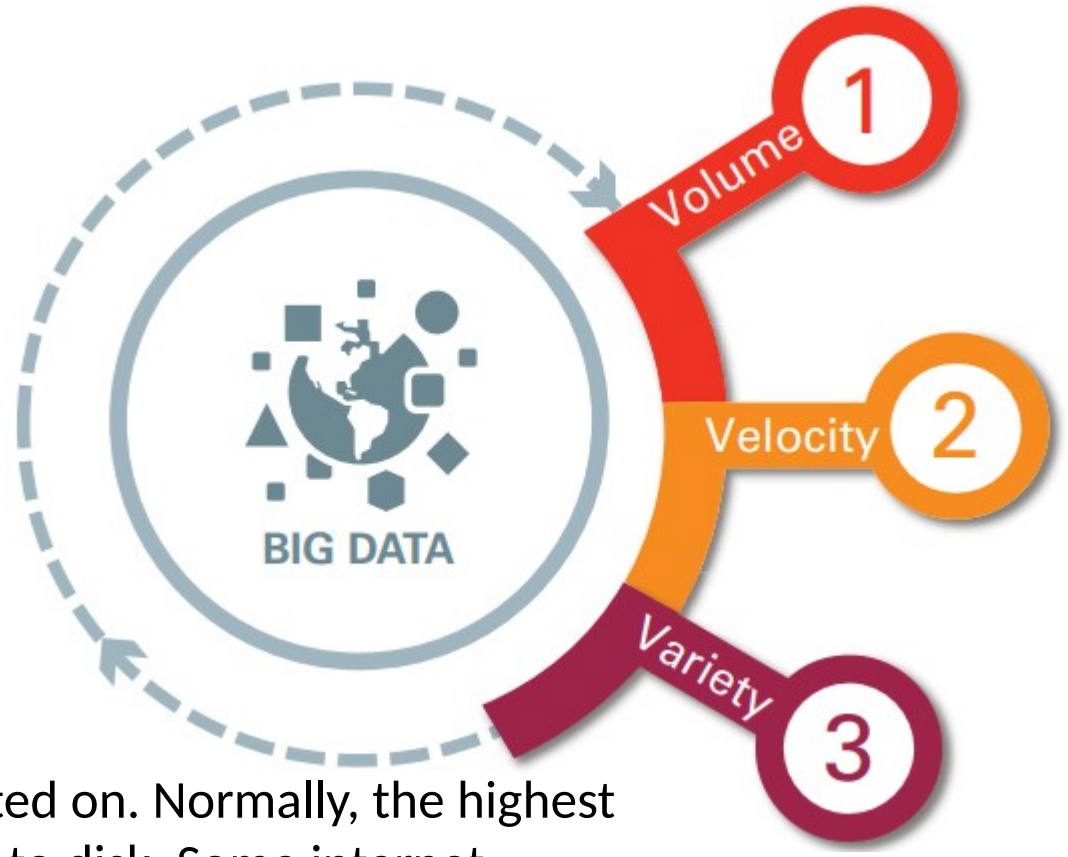
What happened in an internet in a minute



Characteristics of Big Data V3s

❑Volume

The amount of data matters. With big data, you'll have to process high volumes of low-density, unstructured data. This can be data of unknown value, such as Twitter data feeds, click streams on a webpage or a mobile app, or sensor-enabled equipment. For some organizations, this might be tens of terabytes of data. For others, it may be hundreds of petabytes.

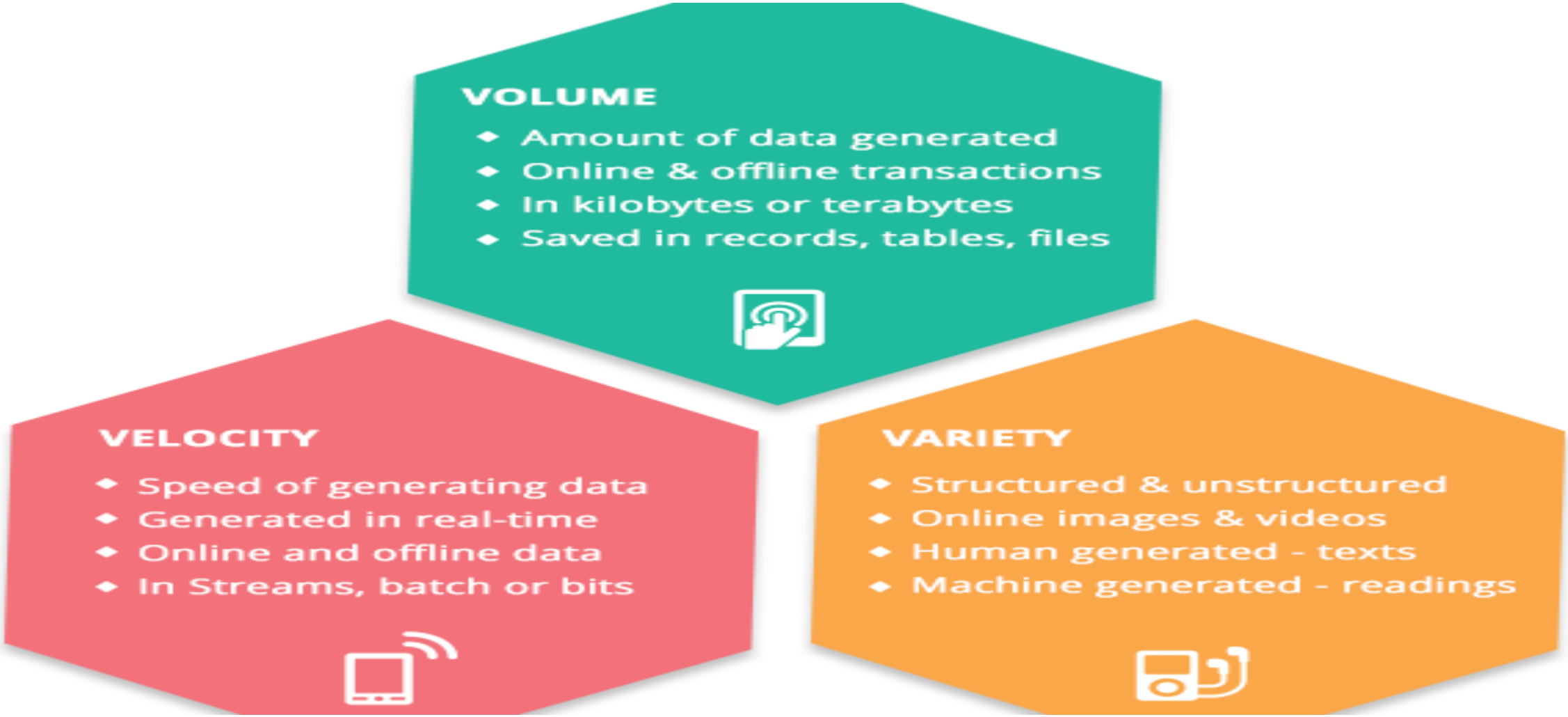


❑ Velocity

Velocity is the fast rate at which data is received and (perhaps) acted on. Normally, the highest velocity of data streams directly into memory versus being written to disk. Some internet-enabled smart products operate in real time or near real time and will require real-time evaluation and action

❑ **Variety**. In today's big data world, data comes in new unstructured data types. Unstructured and semi-structured data types, such as text, audio, and video require addition preprocessing to derive meaning and support metadata.

Characteristics of Big Data V3s…



Big Data Value Chain

- It describe the information flow within a big data system as a series of steps needed to generate value and useful insights from data.
- The Big Data Value Chain identifies the following key high-level activities:

- **Collection –** Structured, unstructured and semi-structured data from multiple sources
- **Ingestion –** loading vast amounts of data onto a single data store
- **Discovery & Cleansing –** understanding format and content; clean up and formatting
- **Integration –** linking, entity extraction, indexing and data fusion
- **Analysis –** Intelligence, statistics, predictive and text analytics, machine learning
- **Delivery –** querying, visualization, real time delivery on enterprise-class availability

Application Of Big Data analytics

Smarter Healthcare

Traffic Control

Manufacturing

Telecom

Trading Analytics

Search Quality

Tools used in Big-Data

- There are number of tools used in BIGDATA. Most popular tools are:

- **Apache Hadoop:** A large data framework is the Apache Hadoop software library. It enables massive data sets to be processed across clusters of computers in a distributed manner. It's one of the most powerful big data technologies, with the ability to grow from a single server to thousands of computers.
- **HPCC :** is a big data tool developed by LexisNexis Risk Solution. It delivers on a single platform, a single architecture and a single programming language for data processing.
- **Apache STORM:** Storm is a free big data open source computation system. It is one of the best big data tools which offers distributed real-time, fault-tolerant processing system. With real-time computation capabilities.
- **Qubole :** Data is Autonomous Big data management platform. It is a big data open-source tool which is self-managed, self-optimizing and allows the data team to focus on business outcomes.
- **Apache Cassandra:** The Apache Cassandra database is widely used today to provide an effective management of large amounts of data.
- **Apache Hive:** Hive is an open-source big data software tool. It allows programmers analyze large data sets on Hadoop. It helps with querying and managing large datasets real fast.

Cont'd...

- **CouchDB:** CouchDB stores data in JSON documents that can be accessed web or query using JavaScript. It offers distributed scaling with fault-tolerant storage. It allows accessing data by defining the Couch Replication Protocol.
- **Pentaho:** Pentaho provides big data tools to extract, prepare and blend data. It offers visualizations and analytics that change the way to run any business. This Big data tool allows turning big data into big insights.
- **Apache Flink:** Apache Flink is one of the best open source data analytics tools for stream processing big data. It is distributed, high-performing, always-available, and accurate data streaming applications.
- **Cloudera** Cloudera is the fastest, easiest and highly secure modern big data platform. It allows anyone to get any data across any environment within single, scalable platform.
- **Open Refine:** OpenRefine is a powerful big data tool. It is a big data analytics software that helps to work with messy data, cleaning it and transforming it from one format into another. It also allows extending it with web services and external data.
- **RapidMiner:** RapidMiner is one of the best open-source data analytics tools. It is used for data prep, machine learning, and model deployment. It offers a suite of products to build new data mining processes and setup predictive analysis.
- **Data cleaner:** Data Cleaner is a data quality analysis application and a solution platform. It has strong data profiling engine. It is extensible and thereby adds data cleansing, transformations, matching, and merging.

Challenges in BIG DATA

- Lack of proper understanding of Big Data
- Data growth issues
- Confusion while Big Data tool selection
- Lack of data professionals
- Securing data
- Integrating data from a variety of sources

NOSQL database

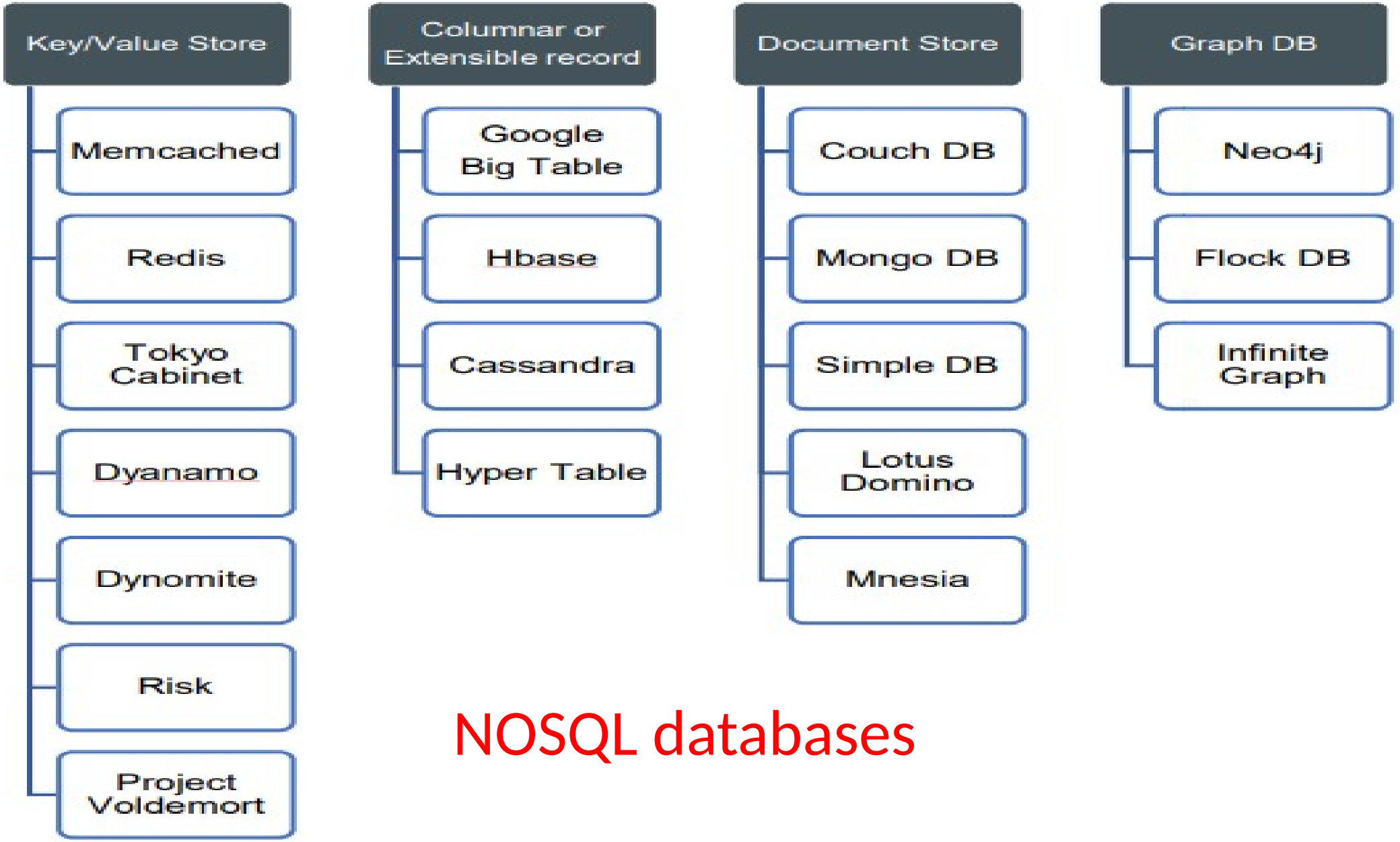
- A NOSQL database is a clever way of cost-effectively organizing large amounts of heterogeneous data for efficient access and updates.
- An ideal NOSQL database is completely aligned with the nature of the problems being solved, and is superfast in accomplishing that task.
- This is achieved by relaxing many of the integrity and redundancy constraints of storing data in relational databases.
- Data is thus stored in many innovative formats closely aligned with business need.
- The diverse NOSQL databases will ultimately collectively evolve into a holistic set of efficient and elegant knowledge stored at the heart of a cosmic computer.

Why NOSQL Database Emerge?

- Big data is, however, a much larger and unpredictable stream of data. Relational databases are inadequate for this task, and will also be very expensive for such large data volumes.
- Managing the costs and speed of managing such large and heterogeneous data streams requires relaxing many of the strict rules and requirements of relational database.
- Depending upon which constraint(s) are relaxed, a different kind of database structure will emerge.
- These are called NOSQL databases, to differentiate them from relational databases that use Structured Query Language (SQL) as the primary means to manipulate dat

Cont’d…

- NOSQL databases are next-generation databases that are non-relational in their design.
- The name NOSQL is meant to differentiate it from antiquated, ‘pre-relational’ databases.
- Today, almost every organization that must gather customer feedback and sentiments to improve their business, uses a NOSQL database.
- NOSQL is useful when an enterprise needs to access, analyze, and utilize massive amounts of either structured or unstructured data that’s stored remotely in virtual servers across the globe.



NOSQL databases

Artificial Intelligence (AI)

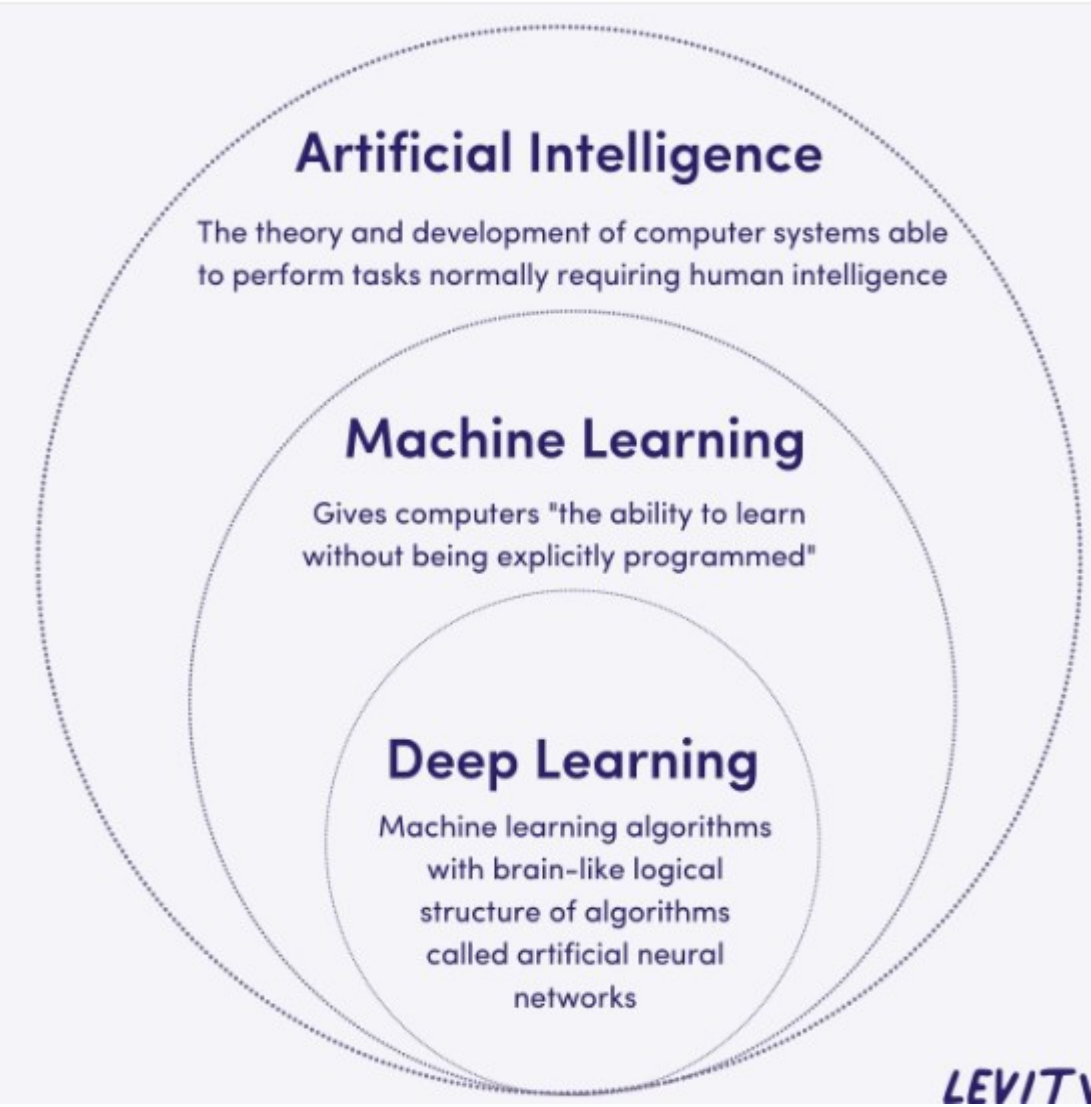
- **Artificial Intelligence (AI)** refers to the development of computer systems or machines that can perform tasks that typically require human intelligence.
- These tasks include understanding natural language, recognizing patterns, solving problems, learning from experience, and making decisions.
- AI systems aim to simulate and replicate human cognitive processes using algorithms, machine learning models, and data analysis.
- *Key progresses of Artificial Intelligence:*
- AI systems designed to perform specific tasks, often better than humans, but within a limited scope.
- Examples:
  - Voice assistants like Siri and Alexa
  - Recommendation algorithms (Netflix, YouTube)
  - Self-driving cars (specific driving tasks)
  - Image recognition software
- A theoretical type of AI that can understand, learn, and apply intelligence to any problem, much like a human being. General AI does not yet exist, but it is a long-term goal in AI research.
- Examples: There are no current examples as General AI remains a future concept.

Subfields of AI:

- **Machine Learning (ML)**:A subset of AI where teaching computers to learn from data and make predictions or decisions without being explicitly programmed.
  - ✓It includes supervised learning, unsupervised learning, and reinforcement learning

  Examples: Spam filters, recommendation systems, fraud detection.
- **Deep Learning: A** subfield of machine learning that uses neural networks with many layers (hence "deep") to model complex patterns in data. It is especially effective in tasks like image recognition, natural language processing, and game playing.

Examples: Face recognition, autonomous driving, advanced natural language understanding models like GPT-4.



Subfields of AI:

- **Natural Language Processing (NLP)**:A branch of AI that focuses on the interaction between computers and human language, enabling machines to read, understand, and generate human language.

  Examples: Chatbots, language translation tools, virtual assistants.

- **Computer Vision**:A subfield of AI that enables machines to interpret and make decisions based on visual data from the world (e.g., images, videos).

  Examples: *Facial recognition, medical image analysis, self-driving car vision.*

- **Robotics:** The integration of AI into robots, allowing them to perform complex tasks in the physical world.

  Examples: Industrial robots, drones, autonomous robots for surgery.

Large Language Models (LLMs)

- Large Language Models (LLMs) are a subset of AI models designed to understand, generate, and process human language.
- These models are built using deep learning techniques, specifically neural networks, and are trained on vast amounts of text data to perform a variety of natural language processing (NLP) tasks, such as translation, summarization, question-answering, and more.
- In addition to teaching human languages to artificial intelligence (AI) applications, large language models can also be trained to perform a variety of tasks like understanding protein structures, writing software code, and more.
- Like the human brain, large language models must be pre-trained and then fine-tuned so that they can solve text classification, question answering, document summarization, and text generation problems.

Large Language Models (LLMs)

- **How large is large?**
- The definition is fuzzy, but "large" has been used to describe BERT (110M parameters) as well as PaLM 2 (up to 340B parameters).
- Parameters are the weights the model learned during training, used to predict the next token in the sequence.
- "Large" can refer either to the number of parameters in the model, or sometimes the number of words in the dataset.

Large Language Models (LLMs)

- **Transformers model**
- A key development in language modeling was the introduction in 2017 of Transformers, an architecture designed around the idea of attention.
- A transformer model processes data by tokenizing the input, then simultaneously conducting mathematical equations to discover relationships between tokens.
- Transformers are the state-of-the-art architecture for a wide variety of language model applications, such as translators.
- If the input is " I am a clever student.", a Transformer-based translator transforms that input into the output " **እኔ ጎበዝ ተማሪ ነኝ**።.", which is the same sentence translated into Amharic.
- Full Transformers consist of an encoder and a decoder.
  - ✓ An **encoder** converts input text into an intermediate representation, and a
  - ✓ **decoder** converts that intermediate representation into useful text.

Large Language Models (LLMs)

- **How do large language models work?**
- A large language model is based on a transformer model and works by
  - receiving an input,
  - encoding it, and then
  - decoding it to produce an output prediction.
- But before a large language model can receive text input and generate an output prediction, it requires training, so that it can fulfill general functions, and fine-tuning, which enables it to perform specific tasks.
- **Training:** Large language models are pre-trained using large textual datasets from sites like Wikipedia, GitHub, or others.
  - ✓ These datasets consist of trillions of words, and their quality will affect the language model's performance.
- **Fine-tuning:** In order for a large language model to perform a specific task, such as translation, it must be fine-tuned to that particular activity.
  - ✓ Fine-tuning optimizes the performance of specific tasks.
- **Prompt-tuning** fulfills a similar function to fine-tuning, whereby it trains a model to perform a specific task through few-shot prompting, or zero-shot prompting they can perform tasks they weren't explicitly trained on.
- For example, given just a few examples, they can follow new instructions or attempt unfamiliar tasks..
  - ✓ *A prompt is an instruction given to an LLM.*

Advantages of large language models:

- **Large language models can be used for several purposes:**
- **Information retrieval:** Think of Google. Whenever you use their search feature, you are relying on a large language model to produce information in response to a query.
  - ✓ It's able to retrieve information, then summarize and communicate the answer in a conversational style.
- **Sentiment analysis:** As applications of natural language processing, large language models enable companies to analyze the sentiment of textual data.
- **Text generation:** LLMs are behind **generative AI**, like ChatGPT, and can generate text based on inputs.
- **Code generation:** Like text generation, code generation is an application of generative AI. LLMs understand patterns, which enables them to generate code.
- **Chatbots and conversational AI:** Large language models enable customer service chatbots or conversational AI to engage with customers, interpret the meaning of their queries or responses, and offer responses in turn.

- **Examples of popular large language models**
- popular LLM models include:

- PaLM: Google's Pathways Language Model (PaLM) is a transformer language model capable of common-sense and arithmetic reasoning, joke explanation, code generation, and translation.
- Bidirectional Encoder Representations from Transformers (BERT) language model was also developed at Google. It is a transformer-based model that can understand natural language and answer questions.
- XLNet: A permutation language model, XLNet generated output predictions in a random order, which distinguishes it from BERT. It assesses the pattern of tokens encoded and then predicts tokens in random order, instead of a sequential order.
- GPT: Generative pre-trained transformers are perhaps the best-known large language models. Developed by OpenAI, GPT is a popular foundational model whose numbered iterations are improvements on their predecessors (GPT-3, GPT-4, etc.).

Generative AI

- Generative AI refers to type of artificial intelligence that can create new content—such as text, images, audio, or video,code—based on the data they've been trained on.
- Unlike traditional AI, which mainly analyzes data or makes predictions, generative AI produces original outputs that resemble the data used for training.
- **Key characterstics of generative AI include:**

- **Text Generation**: Models like GPT (Generative Pre-trained Transformer) that can write human-like text.
- **Image Generation**: Models like DALL·E or Stable Diffusion that can generate images from textual descriptions.
- **Audio/Music Generation**: Models that can generate sounds or compose music, such as Jukedeck or OpenAI's Jukebox.
- **Video Generation**: AI systems that can create new video sequences or even animate characters.
  - Example: **Runway Gen-2** generates video clips based on user input, like "a beach sunset."

Generative AI

Examples of generative AI models and products include:

- ✓ **GPT-4:** OpenAI's flagship generative AI model comes in a variety of sizes.
  It can be accessed through an API or through the major model-hosting platforms
- ✓ **ChatGPT:** An AI language chatbot developed by OpenAI that can answer questions and generate human-like responses from text prompts. It runs on GPT-4.
- ✓ **DALL-E 3:** Another AI model by OpenAI, DALL-E 3 can create images and artwork from text prompts.
- ✓ **Google Gemini:** Previously known as Bard, Gemini is Google's generative AI chatbot and rival to ChatGPT. It's trained on the PaLM large language model and can answer questions and generate text from prompts.
- ✓ **Claude 3.5:** Anthropic's AI model, Claude, offers a 200,000 token context window.
- ✓ **GitHub Copilot:** An AI-powered coding tool, GitHub Copilot suggests code completions within the Visual Studio, Neovim, and JetBrains development environments
- ✓ **Llama 3:** Meta's open-source large language model can be used to create conversational AI models for chatbots and virtual assistants

ChatGPT

- **ChatGPT** is a conversational AI model developed by **OpenAI**, based on the **GPT (Generative Pre-trained Transformer)** architecture.
- It is designed specifically to engage in human-like conversations, answer questions, provide information, and assist users in a wide range of tasks using natural language.

Key Features of ChatGPT:

- **Natural Language Processing**: ChatGPT can understand and generate human-like text based on user input, allowing it to hold conversations, answer questions, and provide recommendations.
- **Generative Model**: Like other GPT models, ChatGPT generates text by predicting the most likely next word in a sentence, which makes it versatile in generating creative content or simulating real dialogue.
- **Pre-trained on Large Datasets**: ChatGPT is pre-trained on vast datasets from books, articles, websites, and other text sources, allowing it to understand language patterns, grammar, facts, and reasoning.
- **Versatility**: ChatGPT can perform a wide range of tasks, including:
  - ✓ Answering questions
  - ✓ Writing essays, code, or creative content
  - ✓ Summarizing long texts
  - ✓ Assisting with tasks like brainstorming ideas or composing emails

Thank You !!!