

DISCUSSION

4V

Volume

1. volume is amount of data or a scale of data per second. Use the terms **terabytes** ;Terabyte: TB = 1,000 GB and **petabytes**; Petabyte: PB = 1,000 TB to discuss the size of data that needs to be processed.
2. The volume of data come from the users which is created, used, and stored from various source such as applications.
3. Big data refers to extremely large and complex data sets that cannot be easily managed or analyzed with traditional data processing tools. Such as scalability, data quality and accuracy.
4. Large and diverse datasets that are huge in volume and also rapidly grow in size over time
5. The huge amount of data can be stored at cloud to provide unlimited storage of data (pay as you go) and to handle disaster recovery. Ex: AWS.
6. Apache airflow, Apache kafka ,Apache Hadoop and MongoDB

Volume

Pibhu Chitburanachart 6580195

Charupat Trakulchang 6481176

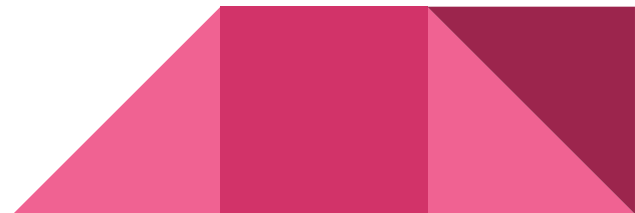
Thitirat Kulpornpaisarn 6580871

Chinanard Sathiseth 6481366

Thananya Osochpromma 6481299

Suttikarn Khuntong 6481305

Tipok Kanngan 6481152



Velocity

- Refers to how fast data can be generated, gathered, and analyzed
 - How quickly data flows into systems and need to be analyzed
- The **speed** is often in real-time or near real-time
- The data trend over time is crucial to help for making predictions or solve problems.
- For example, in medical devices are designed to monitor patients and collect personal data. The data needs to be sent to its destination and analyzed quickly.
- However, in some cases it might be better to have limited set of collected data, since it can be more than could be handled, leading to slower speeds



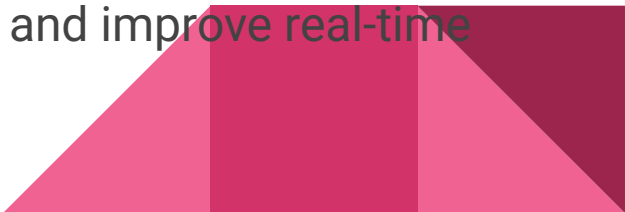
Velocity

- **What is the big challenge?**

It needs to quickly access the wanted data from the database and send it to the user or the target. Otherwise, it also needs to quickly perform adding the input data into the database as well.

- **Solutions**

Implement real-time data processing frameworks such as Apache Kafka, Apache Flink, or Apache Storm. Additionally, using edge computing can help process data closer to the its source, reduce latency, and improve real-time decision-making.



Tools - Velocity

Apache Kafka for real-time data streaming, Apache Storm and Apache Flink for real-time processing, and NoSQL databases like Cassandra and MongoDB for fast data retrieval. Additionally, tools like Apache Spark enable rapid data processing across distributed systems.



Velocity

Thanapoom Tanalakwong 6481205

Promsan Panasakulkan 6481178

Pitchapa Phisutpichet 6580065

Natnicha Sukchuenanant 6580812

Chanon Pluemhathaikij 6581103

Supakorn Panyadee 6581117



Veracity

- Veracity refers to the quality, accuracy, integrity and credibility of your data. It's about ensuring that the data used for analysis and decision-making is reliable and free from errors
- Higher veracity = more meaningful & more important to analyze

Big data can be messy, noisy, and error-prone, which makes it difficult to control the quality and accuracy of the data. Large datasets can be unwieldy and confusing, while smaller datasets could present an incomplete picture. *The higher the veracity of the data, the more trustworthy it is.*

Challenge: Data veracity refers to the biasedness, noise, and abnormality in data. It also refers to incomplete data or errors, outliers, and missing values. *To convert this type of data into a consistent, consolidated, and united source of information creates a big challenge for the enterprise.*



Veracity

- Veracity comes from various factors, including data sources, collection methods, and processing techniques.
- Tools
 - **Apache Spark** : It is an open-source distributed processing solution for big data applications. For quick queries against any data size, it uses in-memory caching and optimized query execution.
 - **RapidMiner** : It is a software package that allows users to perform data mining, text mining, and predictive analytics.



Veracity

- Rapeepat Pokpattanakul 6480358
- Mark Kittiphat Kuprasertwong 6481322
- Pakin Panawattanakul 6580043
- Nitchayanin Thamkunanon 6580081
- Eakawit Nontapot 6481054
- Sirinuttawat Supavachapong 6480963



Variety Group

- Pachara 6480125
- Pavorn 6480138
- Pranai 6481101
- Jitsopin 6480376
- Prompiriya 6480539
- Teetath 6481221
- Norawat 6480566



Variety (What is it)

Different and diverse forms of data that are collected from different sources

- **Structured Data**
 - Pre-formatted tabular data (Ex. SQL, Excel, CSV)
 - Easily organizable and searchable
- **Semi-structured Data**
 - Data without specific formats but contain some organizational properties (Ex. Mail, HTML, XML)
- **Unstructured Data**
 - Data that lacks consistent format (Ex. Text, Audio, Images)

Example: A single company might process text documents, video files, web server logs, and sensor data.



Variety (Where does the data come from)

- User Inputs
 - User interactions with websites (Ex. Tweets, Reels, Posts, Comments)
- Logs from applications and IoT devices
- Metadata & Cookies



Variety (What is the big challenge?)

- The primary challenges of big data variety stem from the difficulty in managing, storing, and processing data from diverse sources and formats. This includes issues with data integration, quality, and the need for specialized tools and platforms to handle the complexity.
- Making sure systems can understand and work with all formats efficiently.



Variety (What solution have been introduced)

- Data Leaks: Allow raw data from different formats to be stored in one place.
- NoSQL Databases: Can handle different data formats
- Data Integration Tools: Help covert, clean, and merge data into a consistent format for analysis.
- Schema on read approaches: Avoid the need to predefined structure; useful for analyzing semi-structured data.
- Machine Learning:



Variety (What tools have been used)

There are 4 tools that are commonly used in Variety

- Apache Hadoop
- Apache Spark
- Tableau
- NoSQL databases (such as MongoDB and Cassandra)

Beyond these there are many more tools and platforms that are use.

