**ASSESSMENT -1 Vinutha S**

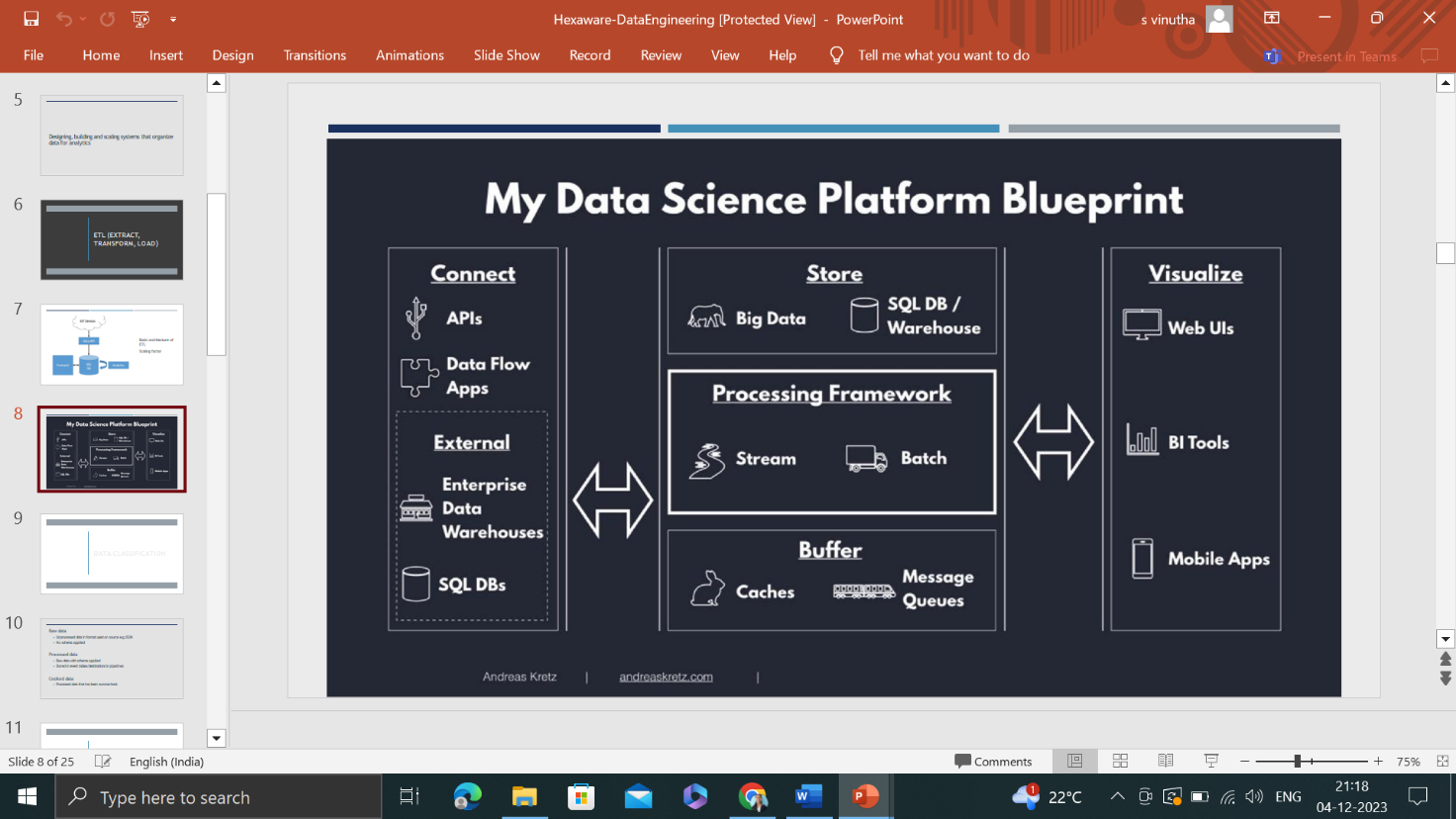
1. **Introduction To Data Engineering**

**Data Engineering vs Data Scientist**

**Data engineering:** Data Engineers focus on building, designing and scaling systems that organize data for analytics

**Data Scientist:** Data Scientist mainly focus on analyzing and interpreting data to find insights and patterns.

1. **Data Science Platform Blueprint**

****

1. **Classification of data:** Data is classified into three types. They are

**Raw Data:** The data which is not processed**.**

**Processed Data:** The data which is processed and schemas applied.

**Cooked data:** The data which is summarized**.**

1. **Processing Framework:** With respect to the data, we do processing.

**Batch processing:** Processing the data in a particular frequency of time.

**Data storage Analytics Insight**

**Stream processing:** The data got processed immediately which is generated from the data sources to the warehouse.

**Ex:** online transaction.

1. **Tools of Big Data**

* Azure data bricks
* Hadoop
* Samza
* Apache Spark

1. **Introduction To Data Warehousing**

**Def:** Data warehousing is the act of organizing and storing data to retrieve efficient and insightful data.

1. **Properties of Data Warehousing**

* **Subject-oriented:** Data are organized according to the subject instead of applications. It mainly focuses on modeling and analysis of data for decision making.
* **Integrated:** Constructed by multiple data sources like relational data base, flat files etc.
* **Time-variant:** Provide information from historical perspective (e g: 5-10 years)
* **Non- volatile:** Once the data enters into the warehouse it will neither be updated nor removed.

1. **Operational data store**

The data that arrived at data warehouse are first passed to Operational Data Store (ODS).

Data is integrated from multiple sources for additional operations on the data.

This integrated data is passed back to operational systems for decision-making

1. **OLTP VS Data warehouse applications**.

OLTP and data warehousing are two important types of data processing systems that are used in a variety of businesses.

OLTP systems are designed to handle high-volume, high-speed transactions.

OLTP systems are used to capture and process transactional data.

Data warehousing systems are designed to store and analyze large amounts of historical data, such as sales data, customer data, and financial data.

data warehousing systems are used to store and analyze that data for business intelligence purposes.

1. **Data marts**

The data in the data warehouse is stored in the form of **Data marts.**

It allows the user to access the data in terms of a specific business line or team.

Data marts are an important tool for business that want to make better decisions based on data.

* **Benefits:** 
  + Improved decision making
  + Increased efficiency
  + Reduce costs
* **Ex: sales D mart**

1. **Data warehouse**

A data warehouse contains data from heterogeneous sources.

It is a central repository where data is processed, transformed, and loaded into the databases so that users can easily access the data from Data Warehouse for analysis using analytical tools for better decision making.

The data may be among these three types:

* Structured data
* Semi structured data
* Unstructured data

1. **Data warehouse life cycle**

Apart from the type of software, life cycles typically include the following phases: requirement analysis, design (including modelling), construction, testing, deployment, operation, maintenance, and retirement.