Data Engineer Profile

Created Date: 2022-08-16

> Metadata
>
> - Title: Section 2: Data Engineer Profile
> - Author: Eshant Garg
> - Reference: https://www.udemy.com/course/dp200exam/

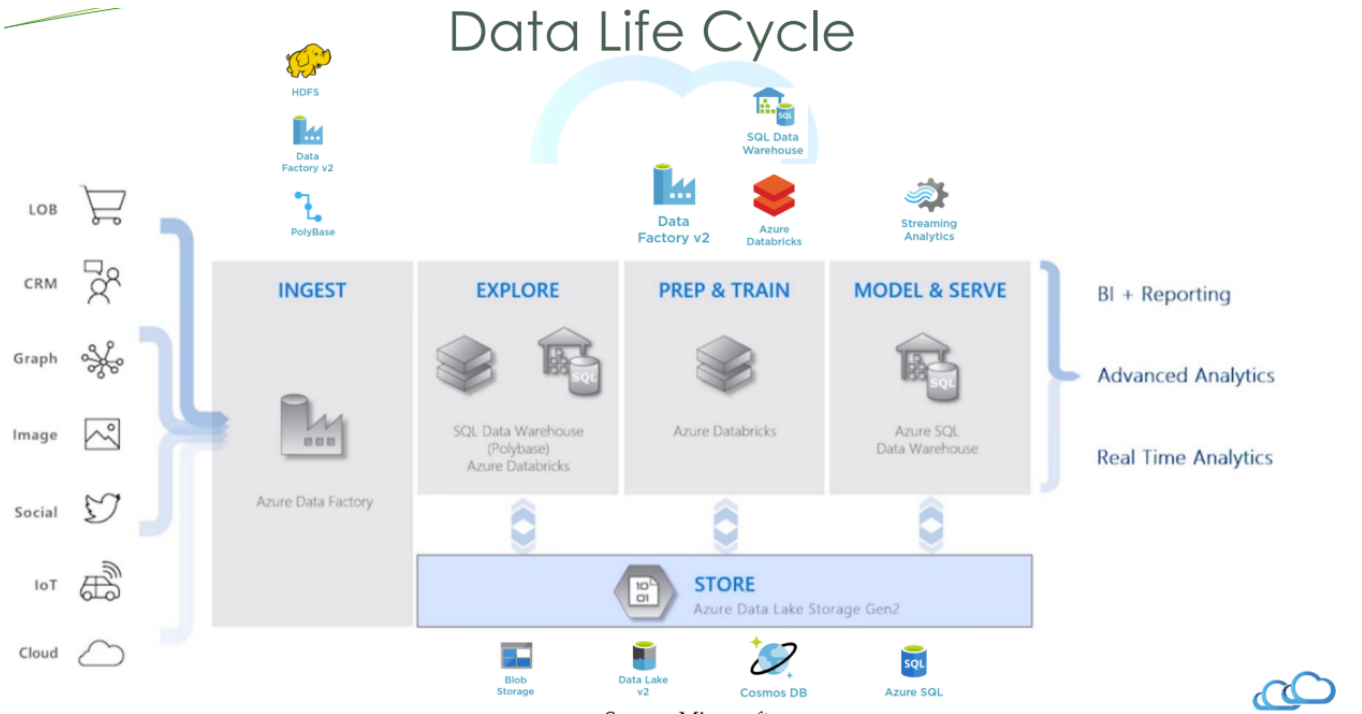
> Links & Tags
>
> - Index: Course Note Index
> - Atomic Tag: #datascience
> - Subatomic Tags: #dataengineering

Data Engineer Roles, Responsibilities, and Technologies

> Understanding the data engineer role requires an understanding of the data life cycle, and how data flows through modern data architecture.

Data Life Cycle

- Ingest
- Explore
- Prep & Train
- Model & Serve



Source: Microsoft

Roles & Responsibilities

- Data Engineers contribute to roughly 80% of the Data Life Cycle, while Data Analyst and Data Scientist mainly contribute to the tail-end of the life cycle (BI/Reporting, Real-Time and Advanced Analytics).
- Data Engineers ensure the smooth flow of data from source to destination, and build the foundation for historic and predictive analytics.
- Key questions for Data Engineer's;
    - What are the sources of data?
    - What is the format of source data?
    - How would I transform this data?
    - What is the destination of this data?
    - What question does this data need to answer?

Data Engineer Technologies

Storage Account

- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data.**
- Suits acting as a **HDInsight Hadoop** data store.

Data Lake Store

- When you need a **low cost, high throughput** data store.
- **Unlimited storage** for **No-SQL** data
- When you **do not need to query** the data directly. **No ad hoc query** support.
- Suits the storage of archive or **relatively static data.**
- Suits acting as a **Databricks** , **HDInsight** and **IoT** data store.

Azure Databricks

- **Eases the deployment** of a Spark based cluster.
- Enables the **fastest processing** of Machine Learning solutions.
- **Enables collaboration** between data engineers and data scientists.
- Provides **tight enterprise security integration** with Azure Active Directory
- **Integration with other Azure Services** and **Power BI.**

Azure CosmosDB

- Provides **global distribution** for both structured and unstructured data stores.
- **Millisecond query response** time.
- **99.999% availability** of data.
- **Worldwide elastic scale** of both the storage and throughput
- **Multiple consistency levels** to control data integrity with concurrency

Azure SQL Database

- When you require a **relational** data store.
- When you need to manage **transactional workloads**
- When you need to manage a **high volume on inserts and reads**
- When you need a service that **requires high concurrency**
- When you require a solution that can scale **elastically**

Azure Data Warehouse

- When you require a **relational** data store.
- When you need to manage **analytical workloads**
- When you need **low cost storage**.
- When you require the ability to **pause and restart the compute**.
- When you require a solution that can scale **elastically**

Azure Stream Analytics

- When you require a **fully managed event processing** engine.
- When you require **temporal analysis of streaming** data.
- Support for analyzing **IoT streaming** data.
- Support for analyzing application data through **Event Hubs.**
- Ease of use with a **Stream Analytics Query Language**.

Azure Data Factory

- When you want to **orchestrate the batch movement** of data.
- When you want to connect to **wide range of data platforms**.
- When you want to **transform or enrich** the data in movement.
- When you want to **integrate with SSIS packages**.
- Enables **verbose logging** of data processing activities.

Azure HDInsight

- When you need a **low cost, high throughput** data store.
- When you need to store **No-SQL** data.
- Provides a Hadoop **Platform as a Service** approach
- Suits acting as a **Hadoop, Hbase, LLAP or Kafka** data store.
- **Eases the deployment and management** of clusters.

Azure Data Catalog

- When you require **documentation** of your data stores.
- When you require a **multi user** approach to documentation.
- When you need to **annotate data sources** with descriptive metadata.
- A **fully managed cloud service** whose users can discover the data sources.
- When you require **a solution that can help business users** understand their data.

Data Engineering Pipeline Processing

- Source: Identify the source systems to extract from.
- Ingest: Identify the technology and method to ingest the data.
- Prepare: Identify the technology and method to transform/prepare data.
- Analyze: Identify the technology and method to analyze the data.
- Consume: Identify the technology to consume/present data.