**Data engineer, data scientist, and artificial intelligence AI engineer**.

**Data engineers** provision and set of data platform technologies that are on-premises and in the Cloud. They manage and secure the flow of structured and unstructured data from multiple sources. The data platforms they use can include relational databases, non-relational databases, data streams, and file stores. Data engineers also ensure that data services integrate securely and seamlessly with other data platform technologies or applications services such as Azure Cognitive Services, Azure Search, or even bots. The Azure Data Engineer focuses on data-related tasks in Azure. Primary responsibilities include using services and tools to ingest, egress, and transform data from multiple sources. Azure data engineers collaborate with business stakeholders to identify and meet data requirements. They design and implement solutions. They also manage, monitor, and ensure the security and privacy of data to satisfy business needs. The role of data engineer is different from the role of a database administrator. A data engineer's scope of work goes well beyond looking after a database and the server where it's hosted. Data engineers must also get, ingest, transform, validate, and clean up data to meet business requirements. This process is called data wrangling. A data engineer adds tremendous value to both business intelligence and data science projects. Data wrangling can consume a lot of time. When the data engineer rankles data, projects move more quickly because data scientists can focus on their own areas of work. Whether your role is a database administrator or business intelligence professional, you can easily transition to a data engineer role. You just need to learn the tools and technology that are used to process large amounts of data.

**Data scientists** perform advanced analytics to extract value from data. Their work can vary from descriptive analytics to predictive analytics. Descriptive analytics evaluate data through a process known as exploratory data analysis, EDA. Predictive analytics are used in machine learning to apply modeling techniques that can detect anomalies or patterns. These are an important part of forecast models. Descriptive and predictive analytics are just one aspect of data scientists' work. Some data scientists might even work in the realms of deep learning. Iteratively experimenting to solve a complex data problem by using customized algorithms. Anecdotal evidence suggests that most of the work in a data scientist project is spent on data wrangling and feature engineering. Data scientists can speed up the experimentation process when data engineers use their skills to successfully wrangle data.

**AI engineers** work with AI services such as cognitive services, cognitive search, and bot framework. Cognitive services include computer vision, text analytics, Bing search, and language understanding, LUIS. Rather than creating models, AI engineers apply the prebuilt capabilities of cognitive services APIs. AI engineers embed these capabilities within a new or existing application or bot. AI engineers rely on the expertise of data engineers to store information that's generated from AI. For example, an AI engineer might be working on a computer vision application that processes images. This AI engineer would ask a data engineer to provision an Azure Cosmos DB instance to store the metadata and tags that the computer vision application generates. The roles of the data engineer, AI engineer, and data scientists differ.

Each role solves a different problem. Data engineers primarily provision data stores. They make sure that massive amounts of data are securely and cost-effectively extracted, loaded, and transformed. AI engineers add the intelligent capabilities of vision, voice, language, and knowledge to applications. To do this, they use the cognitive services offerings that are available out of the box. When a cognitive services application reaches its capacity, AI engineers call on data scientists. Data scientists develop machine learning models and customize components for an AI engineer's application. Each data technology role is distinct and each contributes an important part to digital transformation projects.

Designed data projects in phases that reflect the ELT approach. Source, identify the source systems to extract from. Ingest, identify the technology and method to load the data. Prepare, identify the technology and method to transform or prepare the data. Also, consider the technologies you'll use to analyze and consume the data within the project. These are the next two phases of the process. Analyze, identify the technology and method to analyze the data and consume, identify the technology and method to consume and present the data

**Microsoft Azure Cosmos DB**

Microsoft Azure Cosmos DB is a fully managed NoSQL database for modern app development.

Microsoft Azure Databricks is a data analytics platform optimized for the Microsoft Azure Cloud Services Platform.

Apache Spark is the leading platform for large-scale SQL, batch processing, stream processing, and machine learning

**Azure Stream Analytics**

Azure Stream Analytics to stream and enrich the IoT data to create windows and aggregations, and to integrate Azure Machine Learning. Azure Data Lake Storage Generation 2 to store the bio-metric data at high-speed, Azure Data Factory to perform the extract load transform and final load. Azure Synapse Analytics to provide data warehousing and big data engineering services to support the Chief Medical Officer's needs. Power BI to create the patient dashboard. Part of the dashboard will show real-time telemetry about the patient's condition, the other part will show the patient's recent history, and Azure Machine Learning to process both raw and aggregated data. Researchers will use this to perform predictive analytics on patient re admittance.

The open source world offers four **types of NoSQL databases**. They are

1. key-value store, stores key-value pairs of data in a table structure.
2. Document database stores documents that are tied with metadata to a document searches.
3. Graph database finds relationships between data points by using a structure that's composed of vertices and edges.
4. Column database stores data based on columns rather than rows. Columns can be defined at the queries runtime allowing flexibility in the data that's returned performance.

Azure storage offers four **configuration options**.

1. Azure blob, a scalable object store for text and binary data.
2. Azure files, managed file shares for cloud or on premises deployments.
3. Azure queue, a messaging store for reliable messaging between application components.
4. Azure table and no sequel store for no schema storage of structured data

**Blob storage**

Blob storage works well with images and unstructured data, and it's the cheapest way to store data in Azure. It also provides rest API and SDK for Azure storage in various languages. And supported code languages include .net, java, node, python PHP ruby and go. Azure storage also supports scripting in Azure Power Shell and in the Azure command line interface. Data ingestion, to ingest data into your system use Azure Data factory storage Explorer, the AzCopy tool, Power Shell or visual studio

If you create a storage account as a blob store, you can't query the data directly. To directly query the data, either move the data to a store that supports queries or set up the Azure storage account for a data lake storage

**Azure Data Lake Storage**

Azure Data Lake Storage is a Hadoop compatible data repository that can store any size or type of data.

Data Lake Storage is designed to store massive amounts of data for big data analytics.

Data Lake Storage Gen 2 reduces computation times, making the research faster and less expensive

key features of Data Lake Storage,

1. unlimited scalability,
2. Hadoop compatibility,
3. security support for access-control lists or ACLs,
4. POSIX compliance and optimized Azure Blob Filesystem,
5. ABFS, driver that's designed for big data analytics,
6. zone redundant storage and
7. geo-redundant storage.

Data Lake Storage supports Azure Active Directory ACLs, security administrators can control data access by using the familiar active directory security groups.

**AZURE COSMOS DB**

it is globally distributed multi model database. can deploy databases by using several API models. SQL API, MongoDB API, Cassandra API, Gremlin API, Table API.

1. Mongo DB == semi structured data
2. Cassandra == white columns
3. Gremlin == graph databases.

Azure Cosmos DB supports data encryption type, firewall configurations and access from virtual networks. Data is encrypted automatically user authentication is based on tokens, and Azure Active Directory provides role-based security. Azure Cosmos DB meets many security compliance certifications, including hip hop, fed ramp, SOCS and high trust.

Azure SQL Database helps your application meet security and compliance requirements with a range of built-in features such as advanced threat protection, SQL database auditing, data encryption, Azure Active Directory authentication, multi-Factor authentication, and compliance certification.

**Azure Synapse Analytics**

Azure Synapse Analytics is a Cloud-based data platform that brings together enterprise data warehousing and big data analytics.

key features of Azure Synapse Analytics

1. Quickly run queries at scale: = SQL pools uses massively parallel processing or MPP to quickly run queries across petabytes of data. Because the storage is separated from the compute nodes, you can scale the compute nodes independently to meet any demand at any time.
2. Use Replicated tables for higher performance: = the Data Movement Service or DMS, coordinates and transports data between compute nodes as necessary. But you can use a replicated table to reduce data movement and improve performance.
3. the distributed tables to tune performance: = Azure Synapse Analytics supports three types of distributed tables: hash, round-robin, and replicated. Use these tables to tune performance. Azure Synapse Analytics can also pause and resume the compute layer. This means you pay only for the computation you use.

**Azure Stream Analytics**

Data engineers use Azure stream analytics to process streaming data and respond to data anomalies in real time.

sources include Azure event hubs. Azure IoT hub and Azure Blob storage

**Azure HDInsight**

Azure HDInsight is a low-cost Cloud solution that provides technologies to help you ingest, process and analyze big data.

It supports batch processing, data warehousing, IoT and data science. It includes Apache Hadoop, Spark, HBase, Kafka, Storm, and Interactive Query.

Hadoop includes Apache Hive, HBase, Spark, and Kafka.

1. Hadoop stores data in a file system or HDFS, and Spark stores data in memory.
2. HBase is a NoSQL database built on Hadoop. It's commonly used for search engines. HBase offers automatic failover,
3. Kafka is an open-source platform that's used to compose data pipelines. It offers message queue functionality which allows users to publish or subscribe to real-time data streams.
4. Storm is a distributed real-time streamlining analytic solution. It supports common programming languages like Java, C-sharp and Python.

**Azure Data Catalog**

It is important to note that the data catalog will soon be replaced by Azure purview. This is a unified data governance service that helps you manage and govern your on premises, multi cloud, and software as a service, SAAS data. And you can also easily create a holistic, up-to-date map of your data landscape. With automated data discovery, sensitive data classification, and end to end data lineage, select the hyperlink to discover more.

**Summery**

Azure Synapse Analytics is a Cloud-based data platform that brings together enterprise data warehousing and big data analytics. Data engineers use  
Azure Stream Analytics to process streaming data and respond to data anomalies in real time.

Azure HDInsight is a low-cost Cloud solution that provides technologies to help you ingest, process, and analyze big data.

Azure Databricks provides one-click setup, streamlined workflows, and an interactive workspace for Spark-based applications.

Azure Data Factory is a cloud integration service. It orchestrates the movement of data between various data stores.

Azure Data Catalog is a fully managed Cloud service, it will soon be replaced by Azure Purview