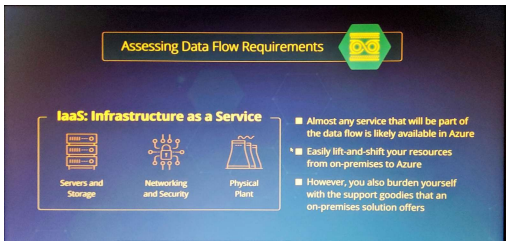
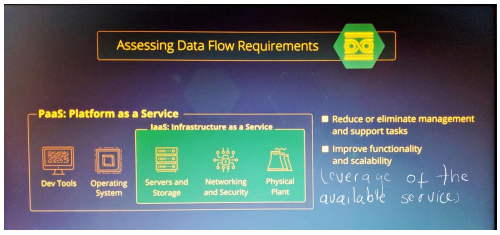
Dataflows2

Monday, 16 January 2023 11:17 AM

Assessing requirements

→ IaaS vs. PaaS





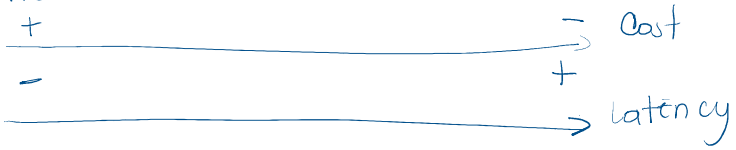
In both cases, it's necessary to ensure the availability and resiliency of it.
HA (High Availability) and DR (Disaster Recovery) How much IaaS and PaaS in play

→ Network bandwith and Network latency. (On Pro ⇄ Cloud) you might consider an Express Route (It offers the ability to mantain a consistent latency)
→ Automated process (scaling, scheduling, triggering)
→ Backups / Monitoring
→ Control flow.
→ Pricing

Azure storage

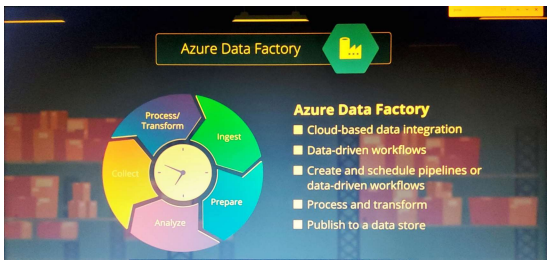
→ Blob storage offers multiple tiers (premium, hot, cool, archive) data storage costs will vary widely, depending on your tiers.

Premium hot cool Archive.



Azure Data Factory

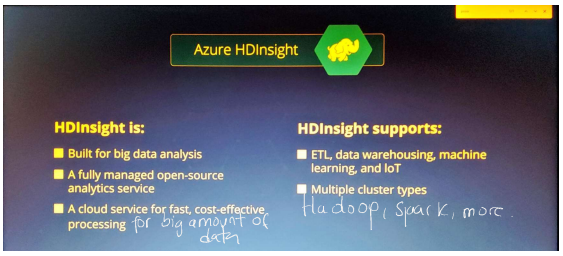
Raw data → store → Process → Info
Azure Data Factory is an Azure offering thats built for complex hybrid ETL, ELT and data integration projects.



Database services

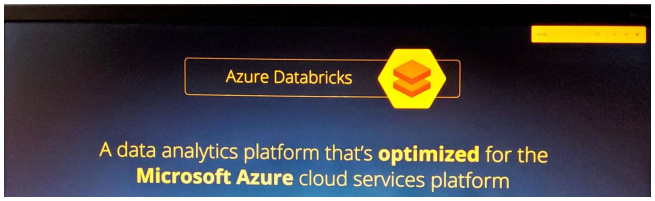
- Azure SQL Database
- Cosmos DB (NO SQL implementation)

Azure HD insights

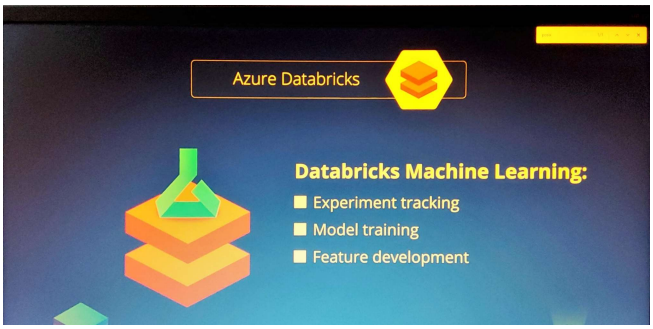


The role of HD Insight in a dataflow is the engine.
It's going to power transformations that need to be completed.
It offers the ability to seamlessly (sin problemas) integrate with many different Azure data stores and services (Cosmos, Data lake)
→ HDInsight can't be paused, you might just want to create an instance of the service. when some sort of analysis or transformation work is required, and then delete it when the work is complete.
→ the programming languages will be dictated by the cluster type.

Azure databricks

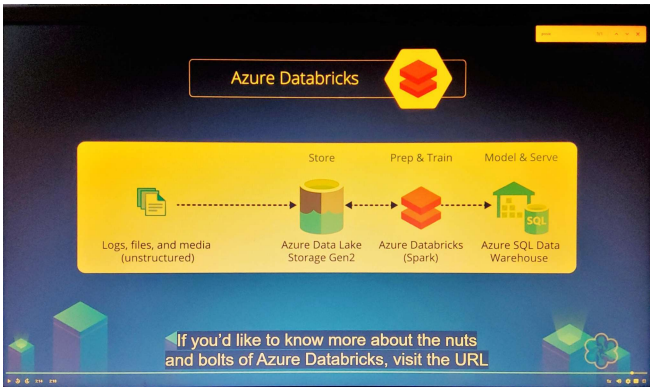


It offers 3 enviroments for developing, data intensive apps, including
(→ DB SQL → to run SQL queries against data lakes, visualizations, dashboards
→ DB Data Science & Engineering → interactive workspace that facilites collaborations
→ DB MLearning.
In Bg Data pipelines, data is tipically ingested into Azure via Azure Data factory
This data is sometimes ingested in batches or streaming via (Kafka, Event Hub, IoT Hub)
the data that is ingested then tipycally winds (Termina) up in Azure Blob Storage or ADLS

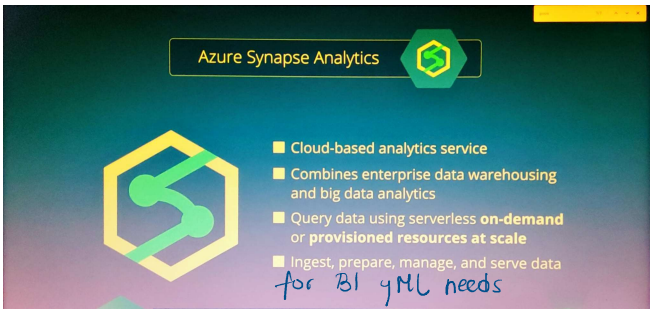


Ultimately, you can use Azure Databricks to you perform ETL operations as part of your [illegible] by extracting data from a source like ADLS Gen 2, pulling into Databrick

Ultimately, you can use Azure Databricks to you perform ETL operations as part of your dataflow by extracting data from a source like ADLS Gen 2, pulling into Databricks, applying transformations in Databricks and the loading the transformed data into Azure Synapse Analytics.

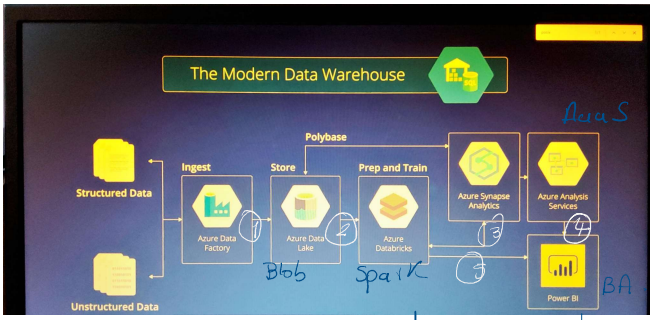


Azure Synapse Analytics



Combines SQL, Spark, Data explores, pipelines for data. ETL-ELT
Also integrates Power BI and Cosmos DB.
→Allows you to import Big Data and to run high-performance analytics on that data,

Modern datawarehouse



Azure Data Factory is a hybrid data integration service, what this does is allow you to create, schedule, and orchestrate ETL and ELT workflows

→The name of your DF is globally unique.