**Building Batch Data Processing Solutions in Microsoft Azure**

[**Course Overview**](https://app.pluralsight.com/player?course=microsoft-azure-batch-data-building&author=tim-warner&name=285745df-beb3-4cdb-ac78-144774a22167&clip=0&mode=live)

[**Course Overview**](https://app.pluralsight.com/player?course=microsoft-azure-batch-data-building&author=tim-warner&name=285745df-beb3-4cdb-ac78-144774a22167&clip=0&mode=live)

[**Developing Batch Processing Solutions with Azure SQL Data Warehouse**](https://app.pluralsight.com/player?course=microsoft-azure-batch-data-building&author=tim-warner&name=3c67ef5b-50a0-454b-b25e-b9b6eeea775b&clip=0&mode=live)

**Overview**

**Preliminary Terminology**

Big Data Definition (Google): Extremely large datasets that may be analyzed computationally to reveal patterns, trends, and associations, especially relating to human behavior and interactions. By BigData, we mean data at scale normally from a variety of sources and in a variety of datatypes.

Concept of 4 vs in big Data nomenclature:

1. Volume: This represents flow of data. It is given in terms of petabytes, exabytes, billions and trillions of records.
2. Velocity: Refers to a processing frequency.
3. Variety: Structured, Unstructured and Semi-Structured.
4. Veracity: Trustworthiness of the data.

There are 3 types of data formats:

1. Structured: This means data that is fully modelled. Ex: SQL Table. Data is divided into number of columns and these columns are all strictly typed and constrained, and relationships among tables are explicitly defined, hence structure.
2. Semi-Structured: Ex: JSON, XML
3. Unstructured: csv, tsv, ordinary text files, images, executables. In Azure Ecosystem these are referred to as Blobs.(Binary Large objects)

**Data warehouse:** Data warehouse is a central repository of integrated data. Purpose of data is already defined. The data is structured and highly transformed, and any operations you perform in the data warehouse are done in a massively parallel way. Basically it’s a very strictly controlled repository of data that you already understand to a large degree.

**Data Lake:** Raw data repository whose purpose is not yet determined and is left in-place until needed. It’s highly accessible and quick to update. A Data Lake is simply a large pool of data that could have come from multiple sources. It's largely unprocessed.

**ETL:** Extract Transform Load. Data is transformed in flight between source and destination. It does not scale very well.

**ELT:** Extract Load Transform. Data is transformed after it’s placed in a data lake. Once on data lake we can perform data transformations and then output your results elsewhere in the data lake or to a relational or non-relational data store, or perhaps out to a reporting platform.

**About Batch Processing**

Batch processing: We work with previously stored data. So we have more tolerance for latency. This normally involves long running batch jobs. Operations include massaging data to prepare for analysis, filtering, aggregating, type casting, changing data, moving/copying data.

Batch Processing workflow with Azure products plugged in : <https://docs.microsoft.com/en-us/azure/architecture/data-guide/big-data/batch-processing>

Data Storage options - Ordinary storage account, Data Lake Storage, SQL Database, Cosmos DB.

Batch Processing - U-SQL, Hive, Pig, and Spark (APIs)

Analytics Data Store – Hbase, Hive, SQL Data Warehouse, Spark SQL

Orchestration – Azure Data Factory, Oozie (HDinsight)

**Data Processing Vs Data Analysis:** Data processing refers to situating data into a usable format. Data analysis is given that cleaned data, how can we ask questions and potentially create visualizations to give us better business intelligence from that data.

Stream processing: We work with incoming data in real time.

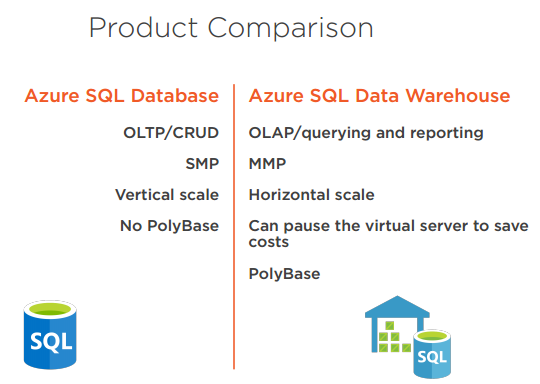
Azure platform has capabilities for both of these processing forms

**Azure SQL DB vs. Azure SQL DW**

* Azure SQL DB is used for OLTP which comprises of create, read, update, delete, CRUD operations, row-level operations, granular queries, row updates, new rows, etc. SQL Data Warehouse is for OLAP. It is used to run detailed and potentially complex queries and reports
* Azure SQL Database uses symmetric multiprocessing where we have a single virtual server that we can scale up or down. Azure SQL Data Warehouse is for massively parallel processing. Under the hood, Azure SQL Data Warehouse has up to 60 or so, compute nodes that we spread our data across.

Azure SQL DW uses horizontal partitioning where we break enormous tables into sections and put them on separate hosts. It's a lot easier to do that in the cloud when we are not provisioning all that hardware yourself.

* With Azure SQL DB, we have vertical Scaling. Azure DW supports horizontal scaling. (Note that we can have Azure SQL DW on premises, but in order to expand compute nodes, we have to go through procurement of hardware and all)
* With Azure DW we can pause the VMs (pause cluster) to save costs. Microsoft provides us this capabilities because SQL DW can be very pricey. (Azure DW is a premier hosted platform for DW)
* Azure DW supports PolyBase. No PolyBase support as of summer 2019 for Azure SQL Database.



The data architecture of Azure DW looks like below:

Data Architecture 
SQL Database 
SQL 
SQL 
Azure TOes 
3rd 
Intelligerxe 
SQL Data Warehouse 
Azure Data Lake 
Database 
Swrces 

Azure DW has the ability to ingest data from variety of data sources in a massively parallel fashion. Data could come from transactional/relational DB, Telemetry streams from Azure tables, Blobs, Azure Data Lake (modification of the Azure storage account that's optimized for parallel processing and big data processes), Hadoop related sources.

As far as taking data out of the warehouse for various purposes, we have integration with the Azure Machine Learning platform, business intelligence stack like Power BI or Excel.

**Demo: Create an Azure Data Lake Storage Gen2 Account**

**Demo: Deploy Azure SQL Data Warehouse**

**Understand Data Lake Storage Gen2**

**PolyBase**

**Demo: Use PolyBase in Azure SQL DW**

**Data Analysis Options**

**Azure Data Factory**

**Demo: Perform ETL Operations with Azure Data Factory**

**For Further Learning**

**Summary**

[**Developing Batch Processing Solutions with Azure HDInsight**](https://app.pluralsight.com/player?course=microsoft-azure-batch-data-building&author=tim-warner&name=59fca865-ecab-4189-83f8-e8a1f623d260&clip=0&mode=live)

**Overview**

**Introducing Apache Hadoop**

**The Hadoop Ecosystem**

**Hadoop vs. Traditional RDBMS**

**Azure HDInsight Architecture**

**Demo: Create an HDInsight Cluster**

**Azure Data Factory/HDInsight Integration**

**Demo: Ingest a Dataset into Data Lake Storage**

**Demo: Perform Data Extraction with Hive**

**Demo: Perform Data Transformation with Hive**

**Demo: Perform Data Loading with Sqoop**

**About Apache Spark**

**About Azure Databricks**

**Demo: Perform Data Visualization with HDInsight and Spark**

**For Further Learning**

**Summary**

[**Developing Batch Processing Solutions with Azure Databricks**](https://app.pluralsight.com/player?course=microsoft-azure-batch-data-building&author=tim-warner&name=bedc4535-c0a9-4b53-9fba-2edc139635ff&clip=0&mode=live)

**Overview**

**Understand Azure Databricks**

**MapReduce vs. Spark**

**The Azure Databricks Ecosystem**

**The Notebook Paradigm in Data Analysis**

**Demo: Deploy an Azure Databricks Service**

**Demo: Define a Cluster and Workspace**

**Demo: Perform ETL with Azure Databricks**

**About Azure Event Hub**

**Demo: Data Processing with Event Hub and Azure Databricks**

**About Azure Batch**

**Azure Distributed Data Engineering Toolkit**

**For Further Learning**

**Summary**