Microsoft Azure Virtual Training Day: Delivering the modern data warehouse
Delivering a modern data warehouse

Nicholas Moore
Cloud Solutions Architect
Agenda

- Why Modern Data Warehousing?
- Building the Modern Datawarehouse
- Advanced Analytics Patterns
- The evolution of Cloud Scale Analytics
Why modern data warehousing?
91% of business leaders see Digital Transformation as a way of sparking innovation and finding efficiencies.

68% say Digital Transformation is increasing profits.

85% say they must offer digital services or become irrelevant.

64% say they have less than 4 years to complete a Digital Transformation or they may go out of business.
Common challenges with on-premises solutions

- Data Silos
- Performance Constraints
- Solution Complexity
- Escalating Costs
Derive real value from your data in the cloud

One hub for all data

Unlimited data scale

Common Platform

Lower TCO
Common customer use cases

Modern data warehouse

“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”

Advanced analytics

“Predict next best offer and customer churn”

Real-time analytics

“Derive insights from our devices and data streams in real-time”
Building the modern data warehouse
Modern data warehouse patterns

Modern data warehouse

“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”

Advanced analytics

“Predict next best offer and customer churn”

Real-time analytics

“Derive insights from our devices and data streams in real-time”
Modern data warehousing patterns

**Ingest & Prep**
- Azure Data Factory
  - Code-free data transformation and ingestion from 90+ data integration connectors
- Azure Databricks (Data prep)
  - Up to 10x faster than vanilla Spark

**Model & Serve**
- Azure Synapse Analytics (Data Warehouse)
  - Up to 14x faster and costs 94% less than other cloud providers

**Visualize**
- Power BI
  - Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*

**Store**
- Azure Data Lake Storage
  - High performance data lake available in all 54 Azure regions

*Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*
Ingest and Prep

- **Ingest & Prep**
  - **Azure Data Factory**
    - Code-free data transformation and ingestion from 90+ data integration connectors
  - **Azure Databricks**
    - (Data prep)
    - Up to 10x faster than vanilla Spark

- **Store**
  - **Azure Data Lake Storage**
    - High performance data lake available in all 54 Azure regions

- **Model & Serve**
  - **Azure Synapse Analytics**
    - (Data Warehouse)
    - Up to 14x faster and costs 94% less than other cloud providers
  - **Power BI**
    - Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*
Model and Serve

Ingest & Prep
- Azure Data Factory
  - Code-free data transformation and ingestion from 90+ data integration connectors
- Azure Databricks
  - (Data prep)
  - Up to 10x faster than vanilla Spark

Model & Serve
- Azure Synapse Analytics
  - (Data Warehouse)
  - Up to 14x faster and costs 94% less than other cloud providers

Visualize
- Power BI
  - Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*

Store
- Azure Data Lake Storage
  - High performance data lake available in all 54 Azure regions

Logs (unstructured)
Media (unstructured)
Files (unstructured)
Business/custom apps (structured)
Visualize

Ingest & Prep
- Azure Data Factory
  Code-free data transformation and ingestion from 90+ data integration connectors
- Azure Databricks (Data prep)
  Up to 10x faster than vanilla Spark

Model & Serve
- Azure Synapse Analytics (Data Warehouse)
  Up to 14x faster and costs 94% less than other cloud providers

Visualize
- Power BI
  Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*

Store
- Azure Data Lake Storage
  High performance data lake available in all 54 Azure regions
Advanced Analytics patterns
The evolving world of Analytics

- Descriptive
- Diagnostic
- Predictive
- Prescriptive
- Cognitive
Advanced Analytics patterns

Modern data warehouse

"Integrate all our data—including Big Data—with our data warehouse for analytics and reporting"

Advanced analytics

"Predict next best offer and customer churn"

Real-time analytics

"Derive insights from our devices and data streams in real-time"
Advanced Analytics patterns

Ingest & Prep
- **Logs** (unstructured)
- **Media** (unstructured)
- **Files** (unstructured)
- **Business/custom apps** (structured)

- **Azure Data Factory**
  - Code-free data transformation and ingestion from 90+ data integration connectors
- **Azure Databricks** (Data prep)
  - Up to 10x faster than vanilla Spark

Model & Serve
- **Azure Synapse Analytics** (Data Warehouse)
  - Up to 14x faster and costs 94% less than other cloud providers

Train
- **Azure Databricks** (Machine Learning)
  - Up to 10x faster than vanilla Spark

Visualize
- **Power BI**
  - Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*

Store
- **Azure Data Lake Storage**
  - High performance data lake available in all 54 Azure regions
Real-time analytics patterns

Modern data warehouse

“Integrate all our data—including Big Data—with our data warehouse for analytics and reporting”

Advanced analytics

“Predict next best offer and customer churn”

Real-time analytics

“Derive insights from our devices and data streams in real-time”
Real-time analytics patterns

Ingest & Prep
- Azure Data Factory
  - Code-free data transformation and ingestion from 90+ data integration connectors
- Azure Databricks (Data prep)
  - Up to 10x faster than vanilla Spark

Model & Serve
- Azure Synapse Analytics (Data Warehouse)
  - Up to 14x faster and costs 94% less than other cloud providers
- Azure Databricks (Machine Learning)
  - Up to 10x faster than vanilla Spark

Visualize
- Power BI
  - Leader in the Magic Quadrant for Business Intelligence and Analytics Platforms*

Store
- Azure Data Lake Storage
  - High performance data lake available in all 54 Azure regions

Streaming
- Azure Event Hub

 Logs (unstructured)
- Media (unstructured)
- Files (unstructured)
- Business/custom apps (structured)
- Sensors and IoT
Demo:
Modern Data Warehousing and Cloud Scale Analytics
The evolution of Cloud Scale Analytics
Azure Synapse Analytics
Limitless data warehouse with unmatched time to insights

PREVIEW

QUERY
Unified experience

Synapse Studio

Integration  Management  Monitoring  Security

Analytics Runtimes

SQL On Demand  SQL Provisioned  Spark

STORE
Azure Data Lake Storage

On-premises data
Cloud data
Devices data
SaaS data

Power BI
Azure Machine Learning
Azure Synapse Analytics

- Limitless scale
- Powerful insights
- Unified experience
- Unmatched security
Ingesting data for analytics workloads

Nicholas Moore
Cloud Solutions Architect
Agenda

- What is Azure Data Factory?
- Ingesting data
- Monitoring
What is Azure Data Factory?
Azure Data Factory

A cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
Azure Data Factory process

Connect & Collect

Transform & Enrich

Publish

Monitor
Azure Data Factory Components

- **Linked Service**
  - Data Lake Store
  - Azure Databricks

- **Control Flow (CF)**

- **Integration Runtime (IR)**

- **Parameters**

- **Triggers**

- **Pipeline**

- **Activities**

- **Dataset**
Ingesting data
Data transformation in Azure

Transforming data with Azure Data Factory

ADF Copy Activity
Connect & Collect

Data ingestion
Load flat files into data lake on a schedule
Azure Data Factory

Data storage
Read data from files using DBFS
Load into SQL DW tables
Azurite Storage/Data Lake Store

Data preparation
Extract and transform relational data
Load processed data into tables optimized for analytics
Azure Databricks

Serving
Load processed data into tables optimized for analytics
Azure Synapse Analytics

Applications
Visualize
Power BI Dashboards

Logs, files, and media (unstructured)

Business and custom apps (structured)

Transactionals storage
SQL DB

Applications manage their transactional data directly

Data prep.
Azure Data Factory

Extract and transform relational data
Reads data from a source data store.

Performs serialization/deserialization, compression/decompression, column mapping, and so on. It performs these operations based on the configuration of the input dataset, output dataset, and Copy activity.

Writes data to the sink/destination data store.
Integration Runtime

Azure Integration Runtime

Self-hosted Integration Runtime
Copy files with the Copy Activity

Supported file formats:
- Text
- JSON
- Avro
- ORC
- Parquet

Copy activity can compress and decompress files with the following codecs:
- Gzip
- Deflate
- Bzip2
- ZipDeflate
Transforming and enriching data

Nicholas Moore
Cloud Solutions Architect
What is Azure Data Factory?
Azure Data Factory

A cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
Azure Data Factory process

- Connect & Collect
- Transform & Enrich
- Publish
- Monitor
Azure Data Factory Components

Linked Service
- Data Lake Store
- Azure DataBricks

Control Flow (CF)

Integration Runtime (IR)

Parameters (@)

Triggers

Pipeline

Activities

Dataset
Component dependencies
Transforming data with the ADF Mapping Data Flow
Data transformation in Azure

Transforming data with Azure Data Factory

Data ingestion
- Load flat files into data lake on a schedule
  - Azure Data Factory

Data storage
- Load into SQL DW tables
  - Azure Storage/ Data Lake Store

Data preparation
- Read data from files using DBFS
  - Azure Databricks
- Load processed data into tables optimized for analytics
  - Azure Synapse Analytics

Transactional storage
- Extract and transform relational data
  - Azure Data Factory

Serving
- Load processed data into tables optimized for analytics
  - Azure Synapse Analytics

Data prep.
- Transform & Enrich

Applications
- Logs, files, and media (unstructured)
- Business and custom apps (structured)
- Visualize
  - Power BI Dashboards

Transform & Enrich
- SQL DB
- Azure Data Factory
Methods for transforming in Azure Data Factory

- Compute resources
- SSIS Packages
- Mapping Data Flow
Methods for transforming data in Azure Data Factory

Code free data transformation at scale
Benefits of Mapping Data Flow

Code free data transformation at scale

- Perform data cleansing, transformation, aggregations, etc.
- Enables you to build resilient data flows in a code free environment
- Enable you to focus on building business logic and data transformation
- Underlying infrastructure is provisioned automatically with cloud scale via Spark execution
Using the Mapping Data Flow

Code free data transformation at scale
Starting the Mapping Data Flow

Code free data transformation at scale
Transformation options in the Mapping Data Flow
Triggering and monitoring
Triggering the Mapping Data Flow

Code free data transformation at scale
Demo:
Transforming your data in Azure Data Factory
In Summary:

Transforming Data with Azure Data Factory

- Azure Data Factory (ADF) is a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
- Transforming data can be performed in ADF by orchestrating a compute resource, calling an SSIS package or using the Mapping Data Flow feature.
  - The Mapping Data Flow feature enables code-free data transformation at scale.
  - Enable you to focus on building business logic and data transformation.
  - It is added to an ADF Pipeline, and can be scheduled or triggered.
  - You can monitor the Mapping Data Flow both visually and programatically.
Demo: Ingesting data with Azure Data Factory
Monitoring data ingestion
# Monitoring

## Pipeline Runs

LoadADLSG1Demo | Monitor Pipeline Runs

**Pipelines** / CopyFromAmazonS3ToADLSG1

- Refresh

## Activity Runs

**Pipeline Run ID** d614f808-7b9d-4362-bb8b-a0bddf226d34

<table>
<thead>
<tr>
<th>Activity Name</th>
<th>Activity Type</th>
<th>Actions</th>
<th>Run Start</th>
<th>Duration</th>
<th>Status</th>
<th>Integration Runtime</th>
</tr>
</thead>
<tbody>
<tr>
<td>Copy-copyfroms3</td>
<td>Copy</td>
<td><img src="image" alt="status" /></td>
<td>01/17/2018, 11:12:45 PM</td>
<td>00:04:00</td>
<td>Succeeded</td>
<td>DefaultIntegrationRuntime (East US 2)</td>
</tr>
</tbody>
</table>
In Summary:

Ingesting Data with Azure Data Factory

- Azure Data Factory (ADF) is a cloud-based data integration service that allows you to orchestrate and automate data movement and data transformation.
- Ingesting data can be performed by the ADF Copy Activity.
- The ADF Copy Activity can be used to connect and collect data for ingestion, and to publish data to BI tools and applications.
- Different Integration Runtimes are required for different ingestion scenarios.
- File copy are very efficient using the ADF Copy Activity.
- You can monitor the performance of the ADF Copy Activity both visually and programmatically.
Data Loading Best Practices

Luis Silva
Cloud Solution Architect, Data Platform
Agenda

- What is Azure Synapse Analytics
- Using Polybase to Load Data in a data warehouse
- Data Loading best practices
What is Azure Synapse Analytics?
A **limitless** analytics service with **unmatched time to insight**, that delivers insights from all your data, **across data warehouses and big data analytics systems**, **with blazing speed**
Azure Synapse Analytics

- PaaS
- Elastic Scale
- Big Data
- Pause/Resume
- Separate Storage/Compute
- Workload Management
Data Warehouse Architecture

Control Node

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110

Compute Node
- 01101010101010101011
- 01010111010101010110
- 01101010101010101011
- 01010111010101010110
SQL Pool Scaling

DW2500c (5 compute nodes)
SQL Pool Scaling

DW5000c (10 compute nodes)
Data Warehouse Processes

Provision → Load → Query

Automate workflow via Azure Data Factory
Loading design goals
Loading design goals

➢ Load data efficiently

➢ Load Data non-obtrusively, respecting concurrent queries and loads

➢ Reduce table fragmentation as much as possible

➢ Provide system recovery capabilities in the event of data load failure with minimal impact on concurrent queries

➢ Multiple methods of loading
Data warehousing loading in Azure

Loading data into a data warehouse in Azure Synapse Analytics

Data ingestion
- Load flat files into data lake on a schedule
  - Azure Data Factory

Data storage
- Load into SQL DW tables
  - Azure Synapse Analytics

Data preparation
- Load processed data into tables optimized for analytics
  - Azure Databricks
- Read data from files using DBFS

Transactional storage
- Applications manage their transactional data directly
  - SQL DB

Data prep.
- Extract and transform relational data
  - Azure Data Factory

Serving
- Load processed data into tables optimized for analytics
  - Azure Synapse Analytics
- Load into SQL DW tables
  - Azure Synapse Analytics

Applications
- Visualize
  - Power BI Dashboards

Logs, files, and media (unstructured)

Business and custom apps (structured)
Loading Methods

BCP
File based

SSIS
Heterogenous

PolyBase
File based
PolyBase benefits

Best practices for loading large amount of data

Leverages MPP architecture
PolyBase is designed to leverage the MPP (Massively Parallel Processing) architecture of Azure Synapse Analytics and will therefore load and export data magnitudes faster than any other tool.

Azure Data Factory support
Azure Data Factory also supports PolyBase loads and can achieve similar performance to running PolyBase manually.

Variety of file formats
PolyBase supports a variety of file formats including RC, ORC and Gzip files.
Components of PolyBase

External Data Source

External File Format

External Tables
Loading best practices
Manage your files

Control Node

Compute Node

Compute Node

Compute Node

Compute Node

Compute Node
Reduce concurrent access
Create a dedicated load user account
Manage singleton updates

Compute Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Compute Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Compute Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Compute Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Computed Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Control Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Control Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Control Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110

Control Node
01101010101010101011
01010111010101010110
01101010101010101011
01010111010101010110
Optimize your loads

Staging data, a 2 step process

- Load into SQL DW tables
- Staging Tables
- Production Tables

Azure Synapse Analytics
Create statistics after loading
Improve the query performance for users

Azure Synapse Analytics

Production Tables
Demo:
Loading data into
Azure Synapse Analytics Data Warehouse
Optimizing data warehousing query performance

Luis Silva
Cloud Solution Architect, Data Platform
Agenda

- What is Azure Synapse Analytics
- Maximizing performance
- Query performance tuning
What is Azure Synapse Analytics?
A **limitless** analytics service with **unmatched time to insight**, that delivers insights from all your data, **across data warehouses and big data analytics systems**, **with blazing speed**
Data Warehouse Processes

Provision → Load → Query

Automate workflow via Azure Data Factory
Data warehouse performance in Azure Synapse Analytics

Query performance

Data ingestion
- Load flat files into data lake on a schedule
  - Azure Data Factory

Data storage
- Load into SQL DW tables
  - Azure Storage/Data Lake Store

Data preparation
- Read data from files using DBFS
  - Azure Databricks
- Load processed data into tables optimized for analytics

Serving
- Load processed data into tables optimized for analytics
  - Azure Synapse Analytics

Applications
- Applications manage their transactional data directly
  - SQL DB

Data prep.
- Extract and transform relational data
  - Azure Data Factory

Visualize
- Power BI Dashboards

Logs, files, and media (unstructured)
- Business and custom apps (structured)
Maximizing Performance
Maximizing Query Performance

Table distribution

- Round Robin Tables
- Hash Distributed Tables
- Replicated Tables
Maximizing Query Performance

Round-robin distribution

- Is the default option for newly created tables
- Evenly distributes the data across the available compute nodes in a random manner, giving an even distribution of data across all nodes
- Loading into Round-robin tables is fast
- Queries on Round-robin tables may require more data movement as data is “reshuffled” to organize the data for the query
- Great to use for loading staging tables
Maximizing Query Performance

Hash distribution

- Distributes rows based on the value in the distribution column, using a deterministic hash function to assign each row to one distribution.
- Is designed to achieve high performance for queries that run against large fact tables in a star schema.
- Choosing a good distribution column is important to ensure the hash distribution performs well.
- As a starting point, use on tables that are greater than 2GB in size and has frequent inserts, updates and deleted.
- But don’t choose a volatile column for the hash distributed column.
Maximizing Query Performance

Replicated Table

- A full copy of a table is placed on every single compute node to minimize data movement.

- Works well for dimension tables in a star schema that are less than 2GB in size and are used regularly in queries with simple predicates.

- Should not be used on dimension tables that are updated on a regular basis.

- You can convert existing round-robin tables to replicated tables to take advantage of the feature using a CTAS statement.
Create statistics after loading

Improve the query performance for users

Azure Synapse Analytics

Production Tables
Demo:
Creating distributed tables
Query Performance Tuning
Query Data Store

- Overcomes the 10,000-row limit of DMV’s output
- Pinpoint and fix queries with plan regression
  - View queries which produce multiple plans
  - 7-day retention period
  - Full query text
- A/B Testing with your Azure Synapse Analytics (SQL DW)
- Identify, improve and tune ad hoc queries
  - Top hitting queries for performance tuning
Query Data Store
Dynamic Management Views

Query

Sys.query_store_query
Query_id (PK)

Plan

Sys.query_store_plan
Plan_id (PK)

Runtime stats

Sys.query_store_runtime_stats
Runtime_stats_id (PK)

Query Text

Sys.query_query_text
Query_test_id (PK)

Runtime stats interval

Sys.query_store_runtime_stats_interval
Runtime_stats_interval_id (PK)

VIEW DATABASE STATE permission
DMVs are in UTC time zone
Query execution with Query Data Store

- Flush to disc every 15 minutes seconds
- 10GB is the max storage size
- Retention period is 7 days
- Maximum plans per query is 200
Azure Synapse Analytics recommendations

Recommendation generation (every 24 hours)

Data skew + Replicate tables
Stats
Tempdb
Adaptive Cache

Recommendation API

Azure Advisor Recommendation

You have free Azure Advisor recommendations!

Azure Advisor is a free offering that analyzes your Azure usage and provides recommendations on how you can save money, improve performance, be more secure, and improve reliability of the solutions you already have running in Azure. Learn more
In Summary:

Query Performance

- Select the proper table distribution
- Detect data skew
  - Use Query Data store
  - Consider changing key columns
  - Only as fast as your slowest distribution
- Provision additional adaptive cache capacity
- Reduce tempdb contention
- Create and update statistics
Demo:
Query Performance Tuning
Thank you!