A CLOUD GURU

DP-203 Study Guide

Brian Roehm
TRAINING ARCHITECT

A CLOUD GURU

Section 2 Review:

Data Engineering Crash Course

Brian Roehm
TRAINING ARCHITECT

Data Transformation

Azure Databricks

Azure Synapse Analytics

Introducing:

1 **Azure Blob**

This is a primary storage service in Azure that includes Azure Data Lake.

2 **Azure Data Factory**

Experience the pipelines of Azure in all their glory.

3 **Azure Synapse Analytics**

If you deal in structured data, Azure Synapse Analytics is definitely the way to go.

Introducing:

4 **Azure Stream Analytics**

Service that provides streaming capability and light transformation.

5 **Azure Databricks**

Service that provides ETL, analytics, and machine learning at a massive scale.

Introduction to Data Lake

Structured vs. Unstructured

SQL	VS.	NOSQL
Relational		Non-relational
Fixed schema		Dynamic
Complex queries		Not for complex queries
Vertical scaling		Horizontal scaling

Data Factory

- **Pipeline**
  - Logical grouping of activities
  - Activities perform a task
- **Activity**
  - Processing steps in a pipeline
  - 3 types of activities
    - Data movement
    - Data transformation
    - Control
- **Datasets**
  - Data structures within the data stores
  - Where the data you need for inputs or outputs lives
- **Linked Services**
  - Connection string needed to connect to data

A CLOUD GURU

Introducing...

Azure Synapse Analytics

Hint Hint... **It's SQL**

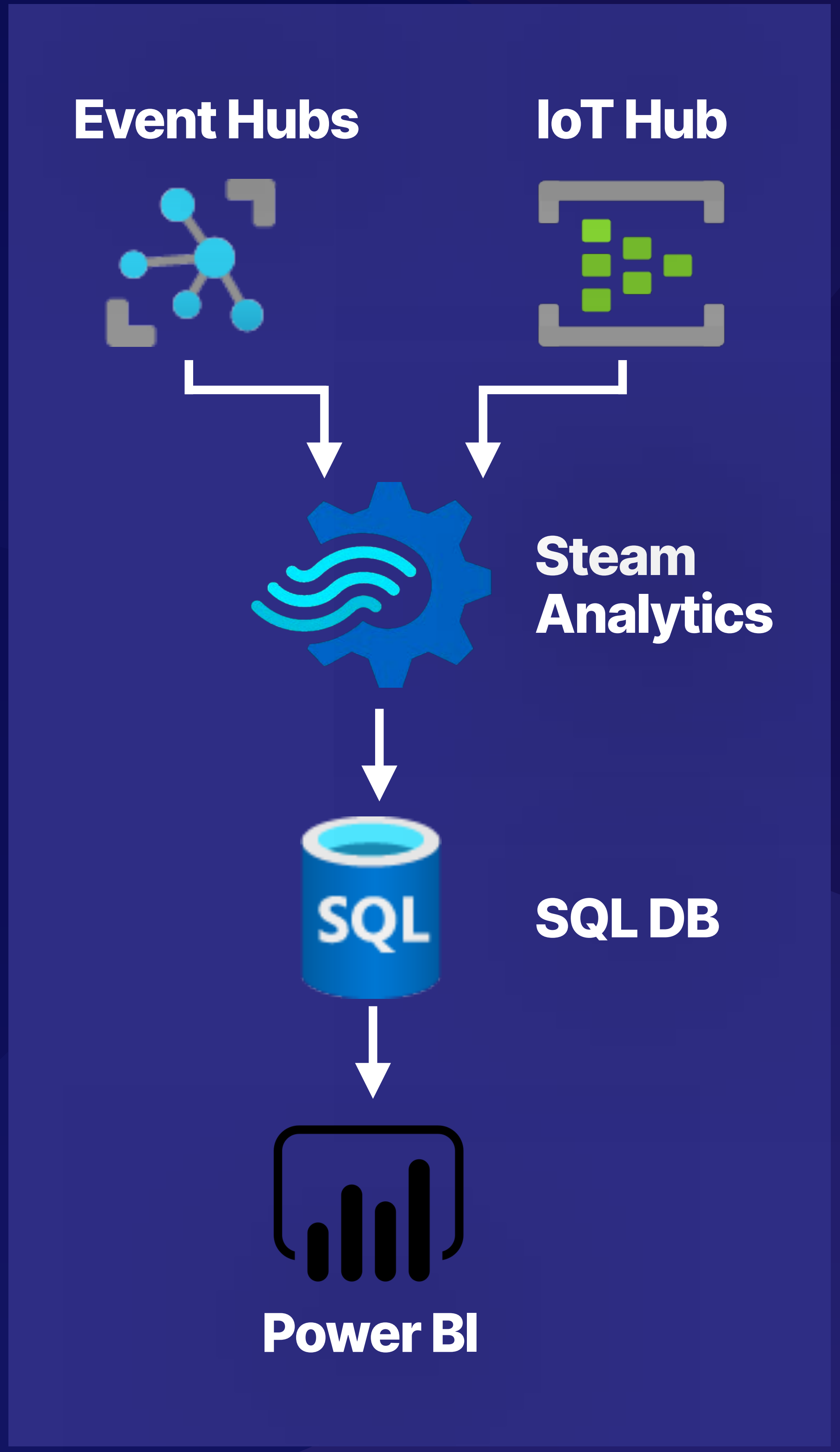
Data Integration, Enterprise Data Warehousing, and Big Data Analytics.

"Bring worlds together with a unified experience to ingest, explore, prepare, manage, and serve data for immediate BI and machine learning needs."

https://azure.microsoft.com/en-us/services/synapse-analytics/#overview

Underwater Closeup of an Azure Synapse Pool

Stream Analytics



Input

- Event Hubs
- IOT Hub
- Blob Storage

Query

- Transformation

Output

- Store and save results

Windowing

- Sliding
- Tumbling
- Hopping

Introducing:

Azure Databricks

3 Main Functions:

- Databricks SQL
- Databricks Data Engineering
- Databricks Machine Learning

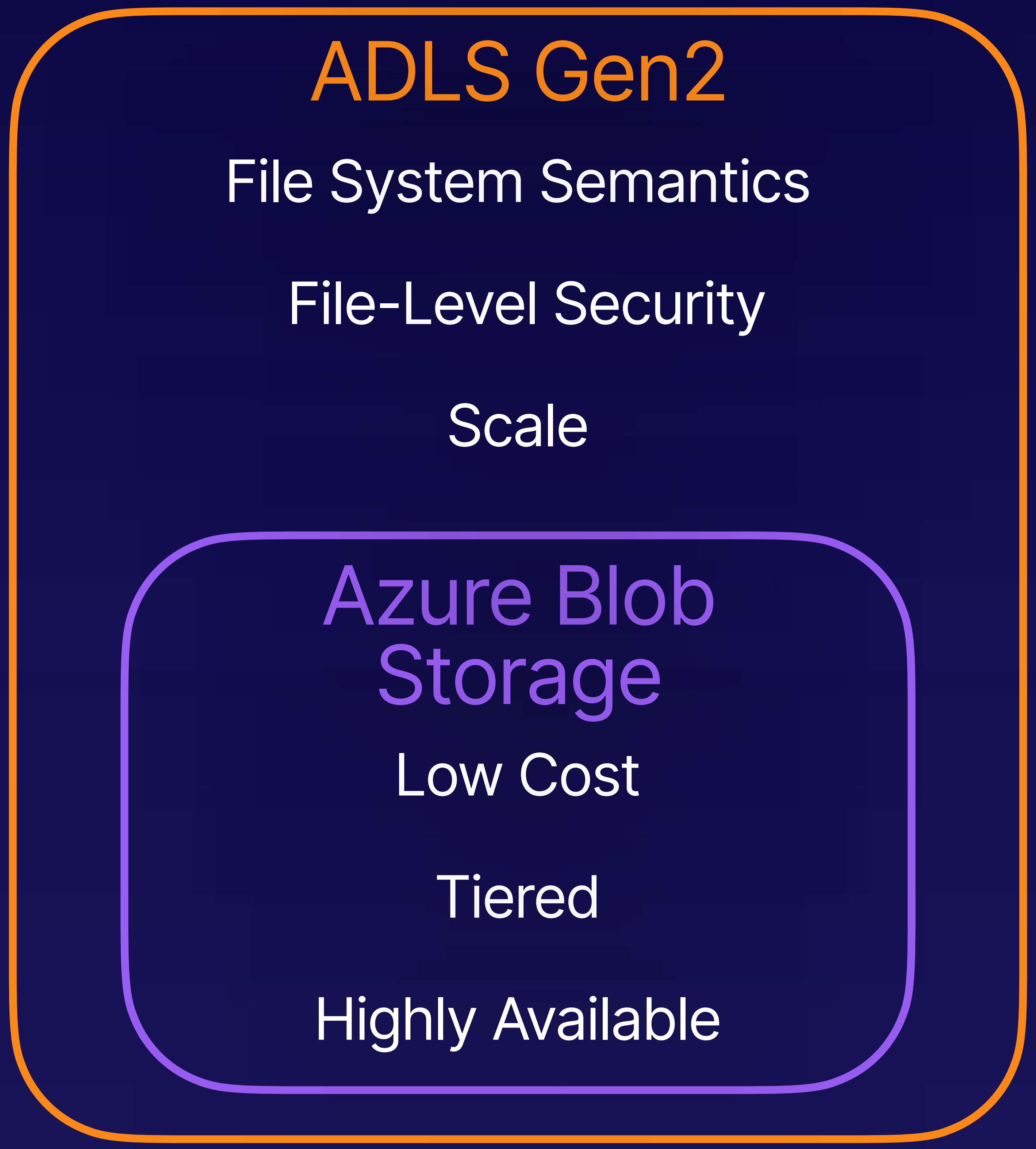
A CLOUD GURU

Section 3 Review: Data Storage

Landon Fowler

TRAINING ARCHITECT

What Is Azure Data Lake Storage Gen2?



ADLS Gen2

File System Semantics

File-Level Security

Scale

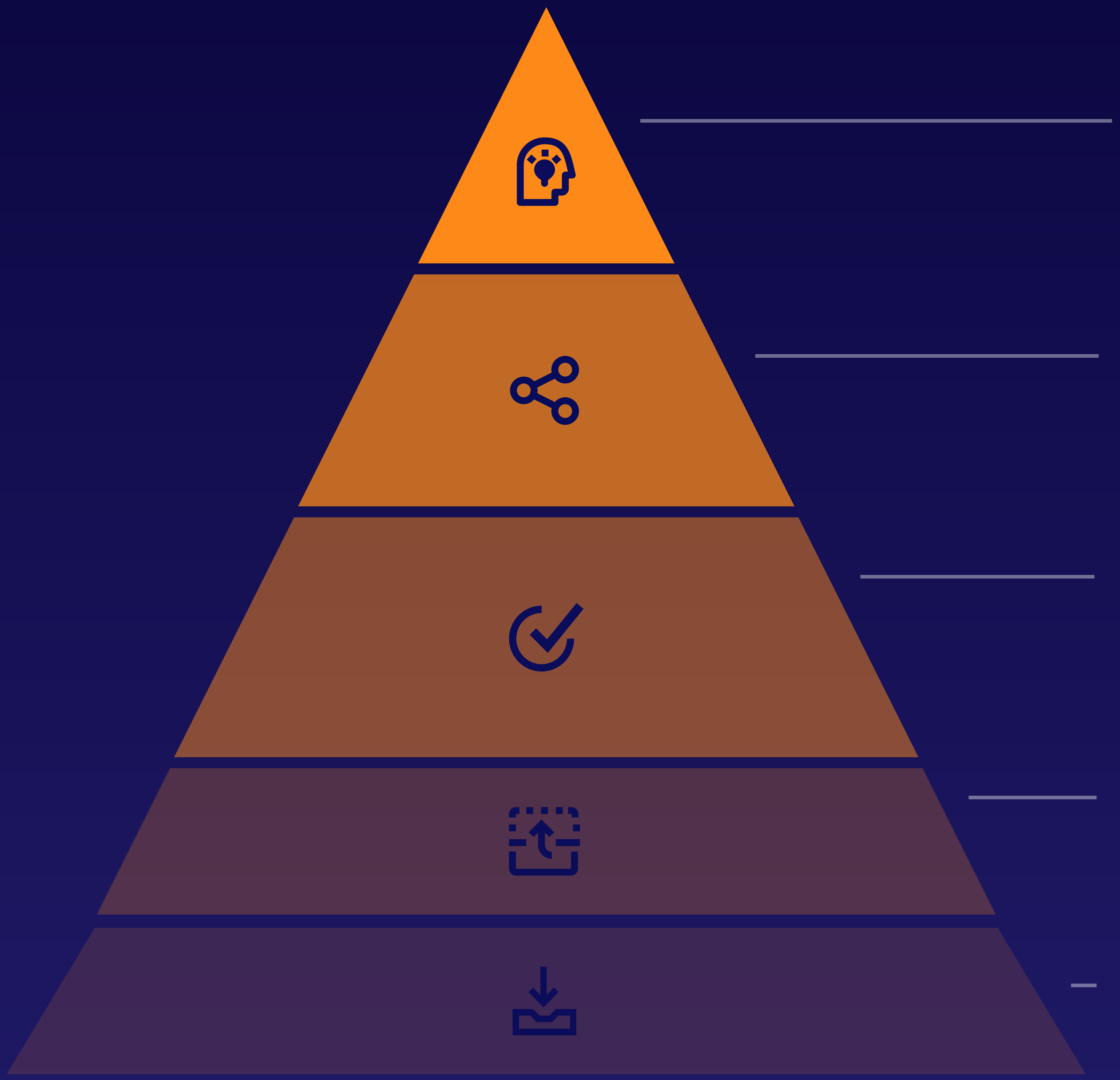
Azure Blob Storage

Low Cost

Tiered

Highly Available

Data Lake Zones



Experimental
A sandbox where data scientists can combine multiple data sets

Production
Easy access point for consumers that includes business logic

Curated
Transformed into consumable data sets

Staging
The first step to refinement by adding basic structure

Landing
Raw data in the original state

Comparison

File Type Showdown

	Avro	Parque	ORC
Analytical Queries		✓	✓
Write Operations	✓		
Nested Data		✓	
ACID Properties			✓
Schema Evolution	✓		

Why It Matters

5 Reasons to Partition

✓ Improve Scalability

✓ Improve Performance

✓ Improve Security

✓ Improve Availability

✓ Improve Cost Savings

Horizontal Partitioning (Sharding)

ID	Name	Topic
1	Amanda	Azure
2	Landon	Data
3	Stosh	DevOps
4	Tia	AWS

A - M

ID	Name	Topic
1	Amanda	Azure
2	Landon	Data

N - Z

ID	Name	Topic
3	Stosh	DevOps
4	Tia	AWS

Vertical Partitioning

ID	Name	Topic	Hours Watched
1	Amanda	Azure	30
2	Landon	Data	60
3	Stosh	DevOps	10
4	Tia	AWS	50

ID	Name	Topic
1	Amanda	Azure
2	Landon	Data
3	Stosh	DevOps
4	Tia	AWS

ID	Hours Watched
1	30
2	60
3	10
4	50

PARTITIONING DATA

Functional Partitioning

ID	Name	Topic
1	Amanda	Azure
2	Landon	Data

ID	Name	Topic
1	Amanda	Azure
2	Landon	Data

ID	Customer	Address
C1	Awesome Co.	123 Corp. Drive
C2	Elite Gurus	456 Guru Way

ID	Customer	Address
C1	Awesome Co.	123 Corp. Drive
C2	Elite Gurus	456 Guru Way

A CLOUD GURU

Partitioning in Azure Data Lake

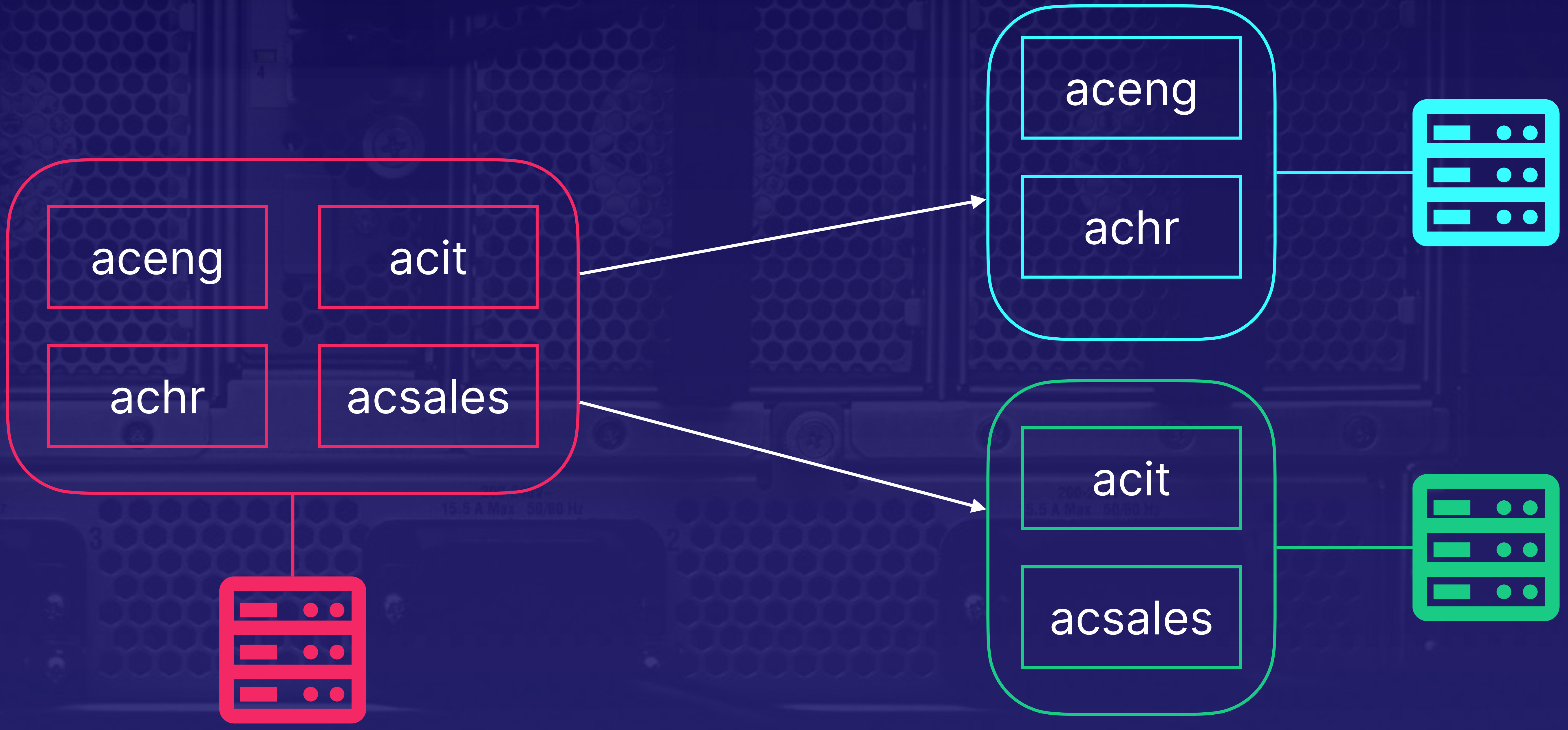
How It Works.

Partition Key

The partition key consists of the full blob name (account + container + blob).

Range-Based Partitioning

The data is split into ranges, which are load-balanced across the storage system.



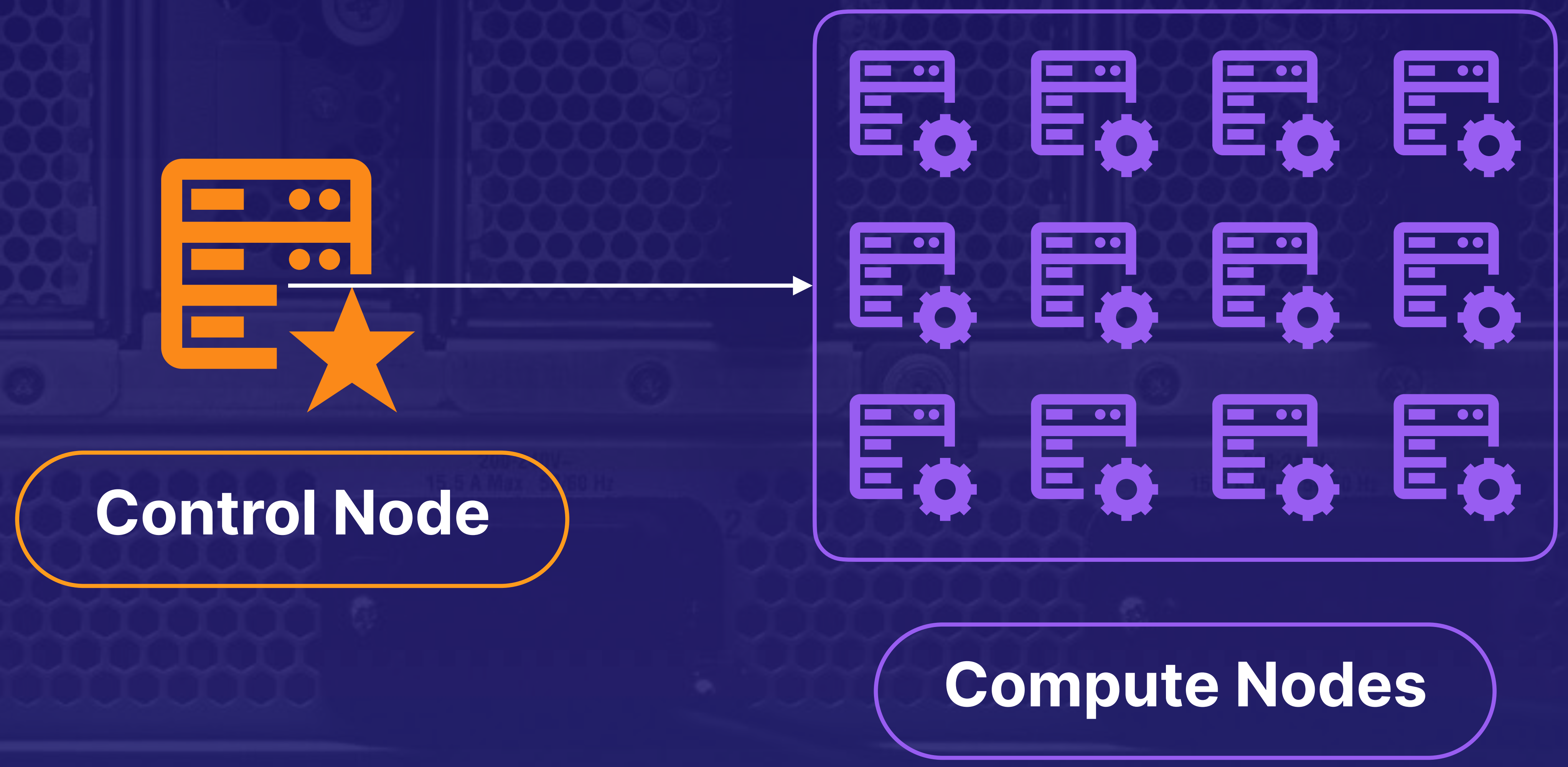
A CLOUD GURU

Partitioning in Azure Synapse

How It Works.

Massively Parallel Processing

Clients connect to a control node, which passes the distributed query to compute nodes. Those execute the work in parallel.



Partitioning in Azure Synapse

How It Works.

Massively Parallel Processing

Clients connect to a control node, which passes the distributed query to compute nodes. Those execute the work in parallel.

Distributed by Default

The data is automatically distributed across 60 underlying databases (distributions).

100 Partitions × 60 Distributions = **6000 Partitions**

A CLOUD GURU

Distribution Types

Round-Robin Distributed

Data is distributed evenly in a random fashion.

Hash Distributed

Data is distributed deterministically by using a hash function.

Replicated

A full copy of the table is replicated to every compute node.

ARCHIVING DATA

Access Tiers

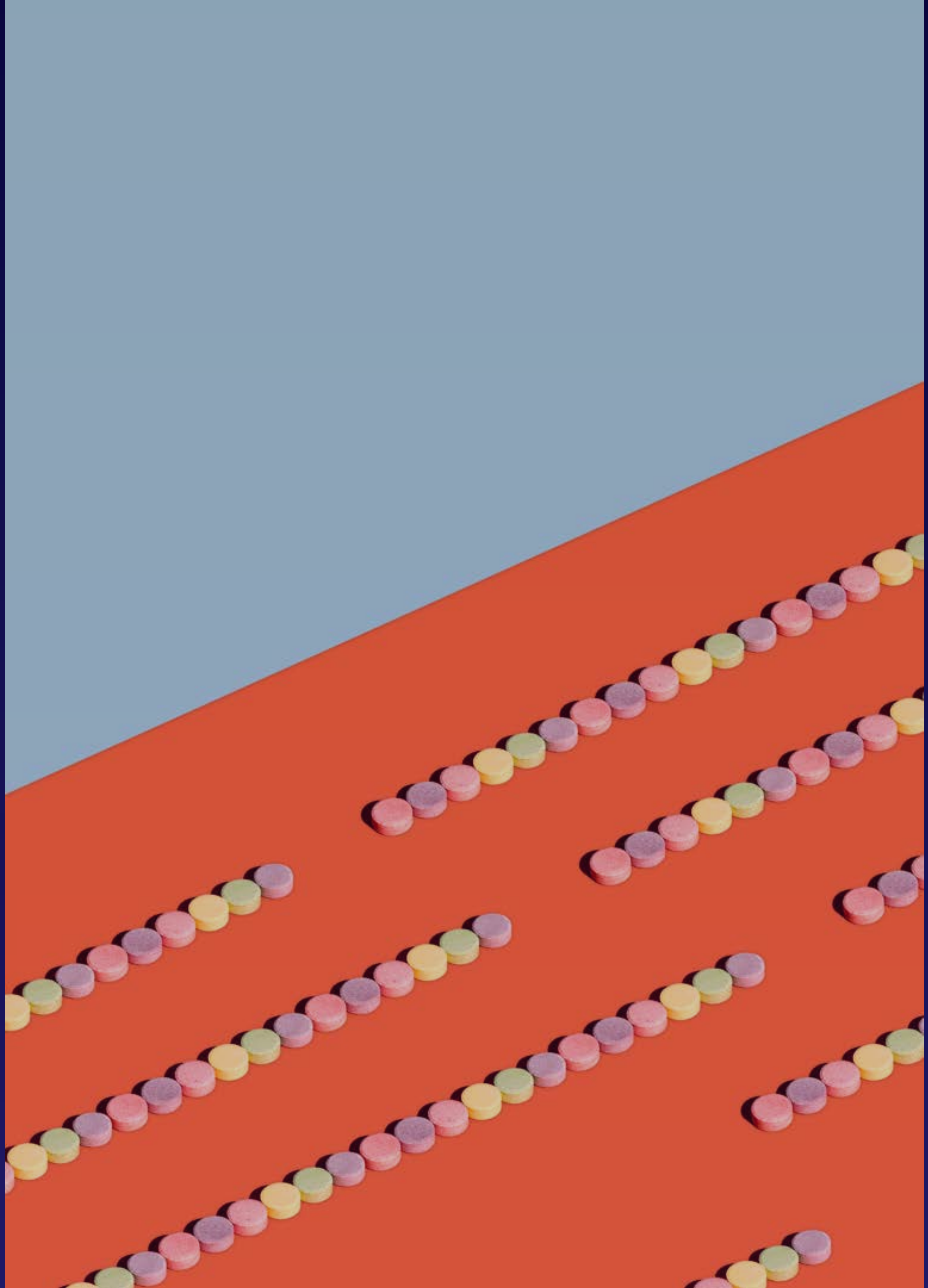
Tier Breakdown

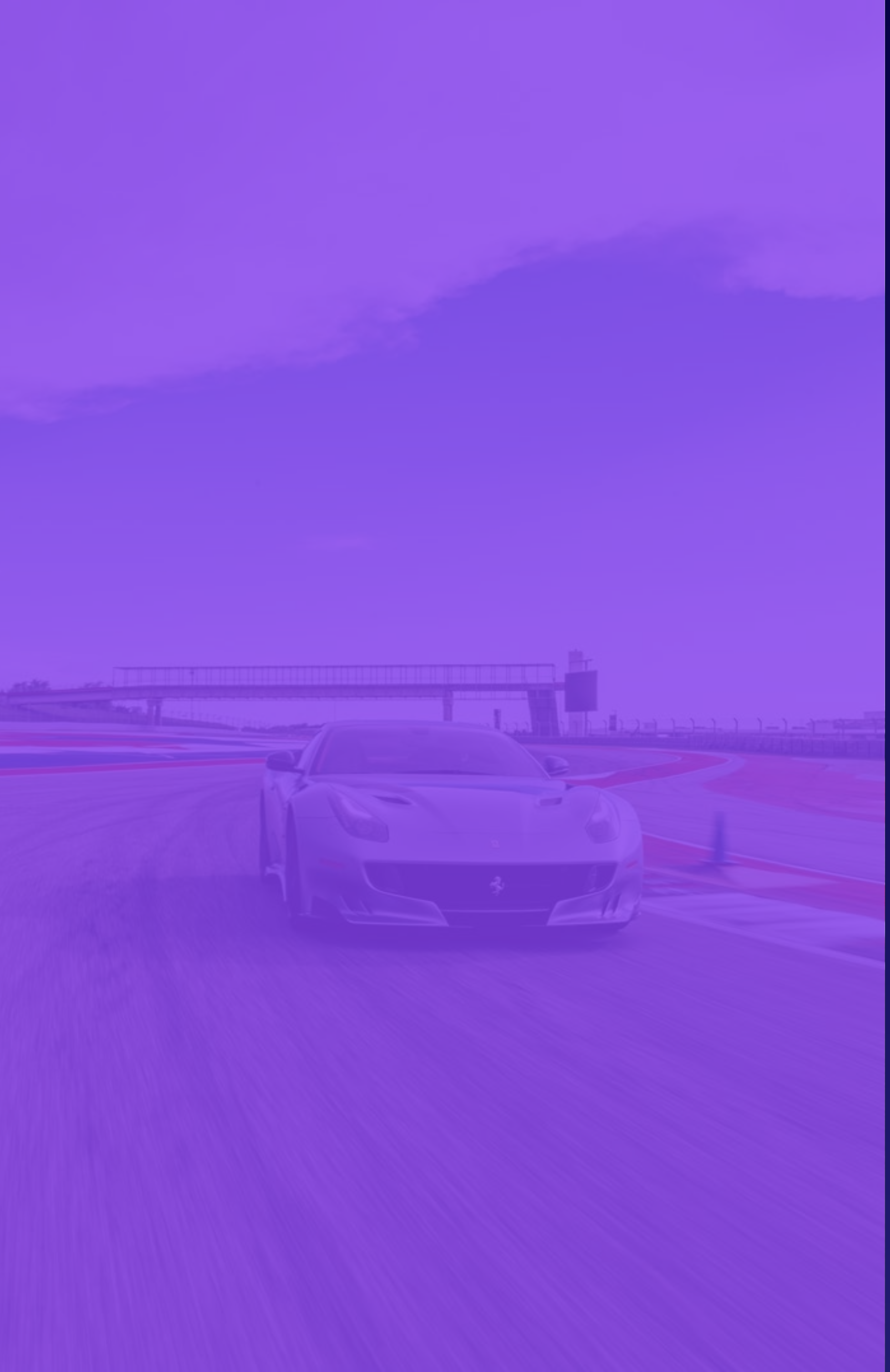
	Minimum Time	Storage Cost	Access Cost	Use
HOT	N/A	↑	↓	Active or staging data
COOL	30 Days	↑ — ↓	↑ — ↓	Short-term data
ARCHIVE	180 Days	↓	↑	Long-term backup

Z-Ordering

A technique for colocating related information in the same set of files.

It is automatically used by the data-skipping algorithms of Delta Lake on Databricks to substantially reduce the amount of data to be read.





Dynamic File Pruning (DFP)

1. **Can Dramatically Improve Query Performance**
   Allows files to be skipped within partitions.

2. **Performance Impact Correlated to Clustering**
   Relies on pre-sorted data, such as Z-Ordering clustering.

3. **Hero of the Non-partitioned**
   DFP is especially efficient for non-partitioned tables, or for joins of non-partitioned columns.

Compressing Data

1. For rowstore objects, you can use row or page compression.

2. Columnstore objects have columnstore compression by default.

3. For additional size reduction on columnstore objects, columnstore archival compression can be enabled.

Sharding Strategies

The Lookup Strategy

The sharding logic uses a map to route requests to the appropriate shard based on the shard key.

The Range Strategy

Related items are grouped together in the same shard and ordered sequentially by the shard key.

The Hash Strategy

A hash of one or more attributes is used to determine the shard in which an item will be placed.

Overview

Always Replicated

Azure Storage creates multiple copies of your data by default.

Ready for Failures

Meet your availability and durability targets by capitalizing on redundancy options.

Have It Your Way

Weigh the tradeoffs between lower costs and higher availability to choose the option right for you.

Primary Region Redundancy

Protecting Home Base

LRS

Locally Redundant Storage

3 synchronous copies within a single physical location.

BONUS TIP

For Azure Data Lake Storage Gen2, Microsoft recommends using ZRS in the primary region.

ZRS

Zone-Redundant Storage

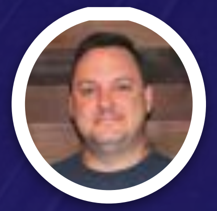
3 synchronous copies across Azure availability zones in this region.

A CLOUD GURU

Section 4 Review: Data Ingestion and Transformation

Landon Fowler

TRAINING ARCHITECT



Brian Roehm

TRAINING ARCHITECT

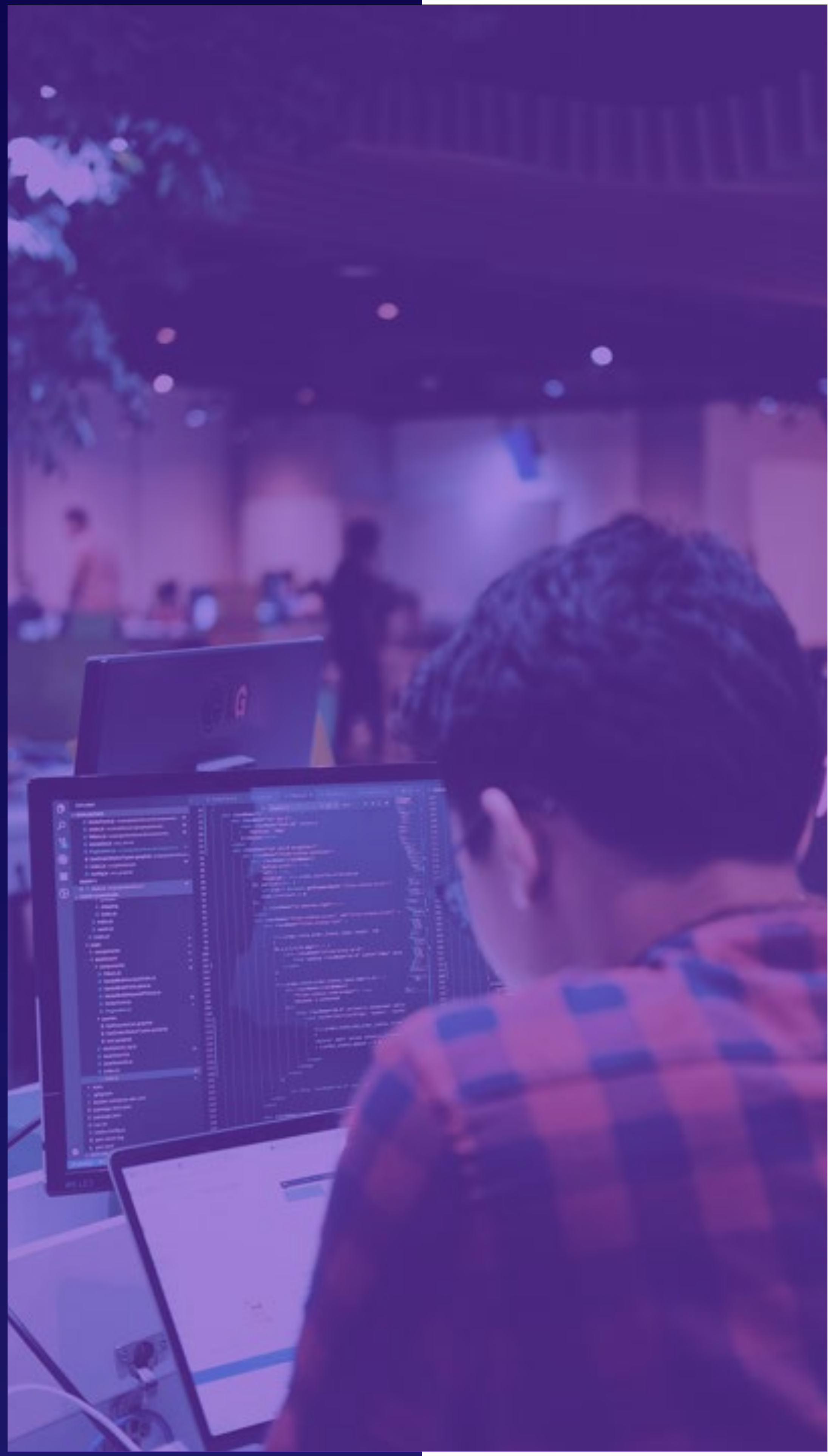
Core Concepts

- **Pipeline**
  - Logical grouping of activities
  - Activities perform a task
- **Activity**
  - Processing steps in a pipeline
  - 3 types of activities
    - Data movement
    - Data transformation
    - Control
- **Datasets**
  - Data structures within the data stores
  - Where the data you need for inputs or outputs lives
- **Linked Services**
  - Connection string needed to connect to data

T-SQL Uses

T-SQL is a powerful language that can be used in a variety of scenarios to move and transform data.

It can be utilized within Azure Machine Learning using the Apply SQL Transformation module.



- Create tables for results or to save datasets.
- Perform custom transformations on data types, or create aggregates.
- Filter or alter data and return the query results as a data table.

Azure Synapse Pipelines and Data Factory

Use For:

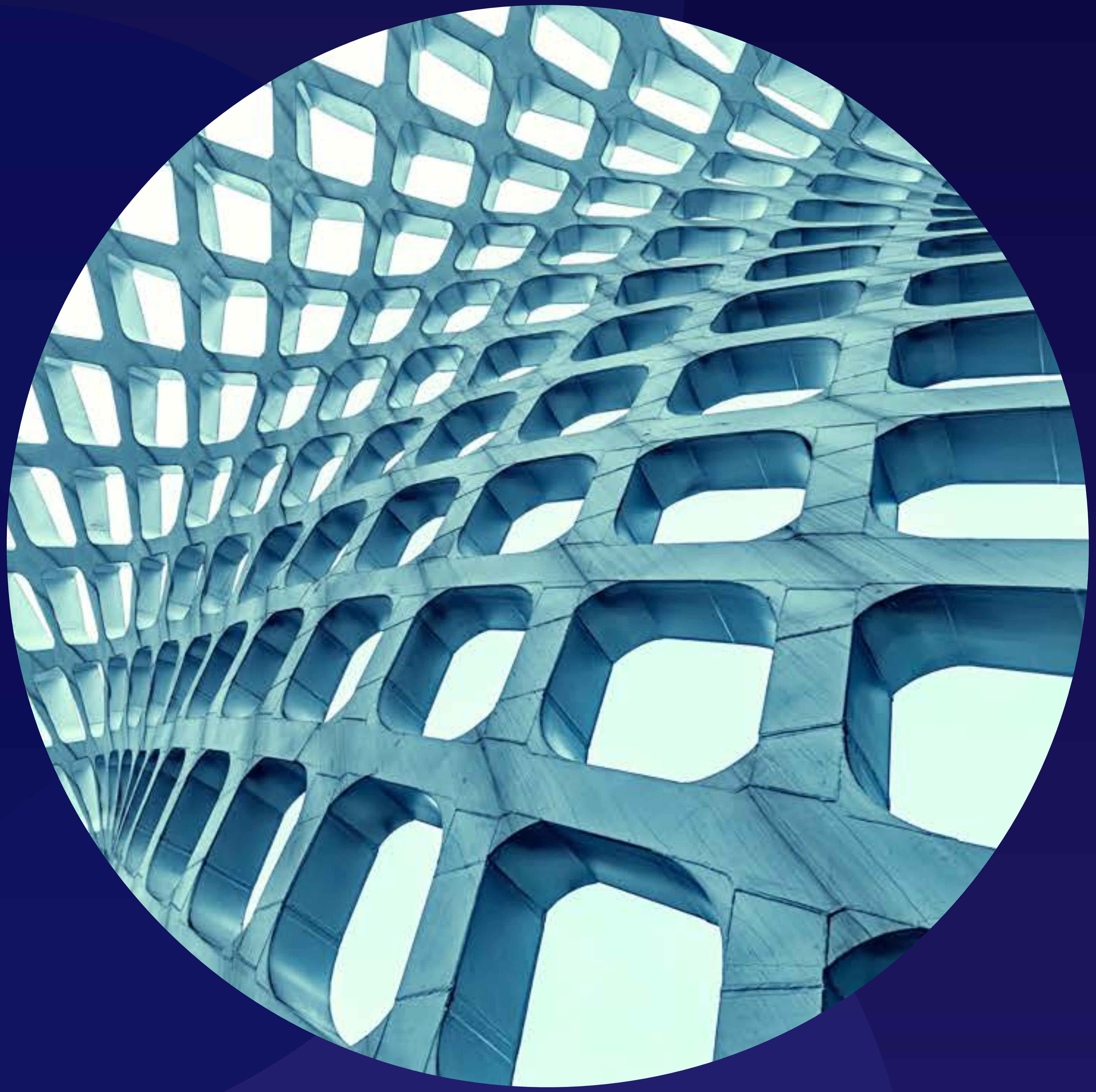
Analytics Projects

When building an analytics solution, Synapse Analytics is a one-stop shop with a fully integrated design experience.

	ADF	Synapse
SSIS Activity	✓	
Power Query Activity	✓	
Monitoring of Spark Jobs for Data Flow		✓
Azure Monitor Integration	✓	

https://docs.microsoft.com/en-us/azure/synapse-analytics/data-integration/concepts-data-factory-differences

Overview



What Is Scala?

A programming language leveraged in Azure Databricks for ETL and data analysis operations.

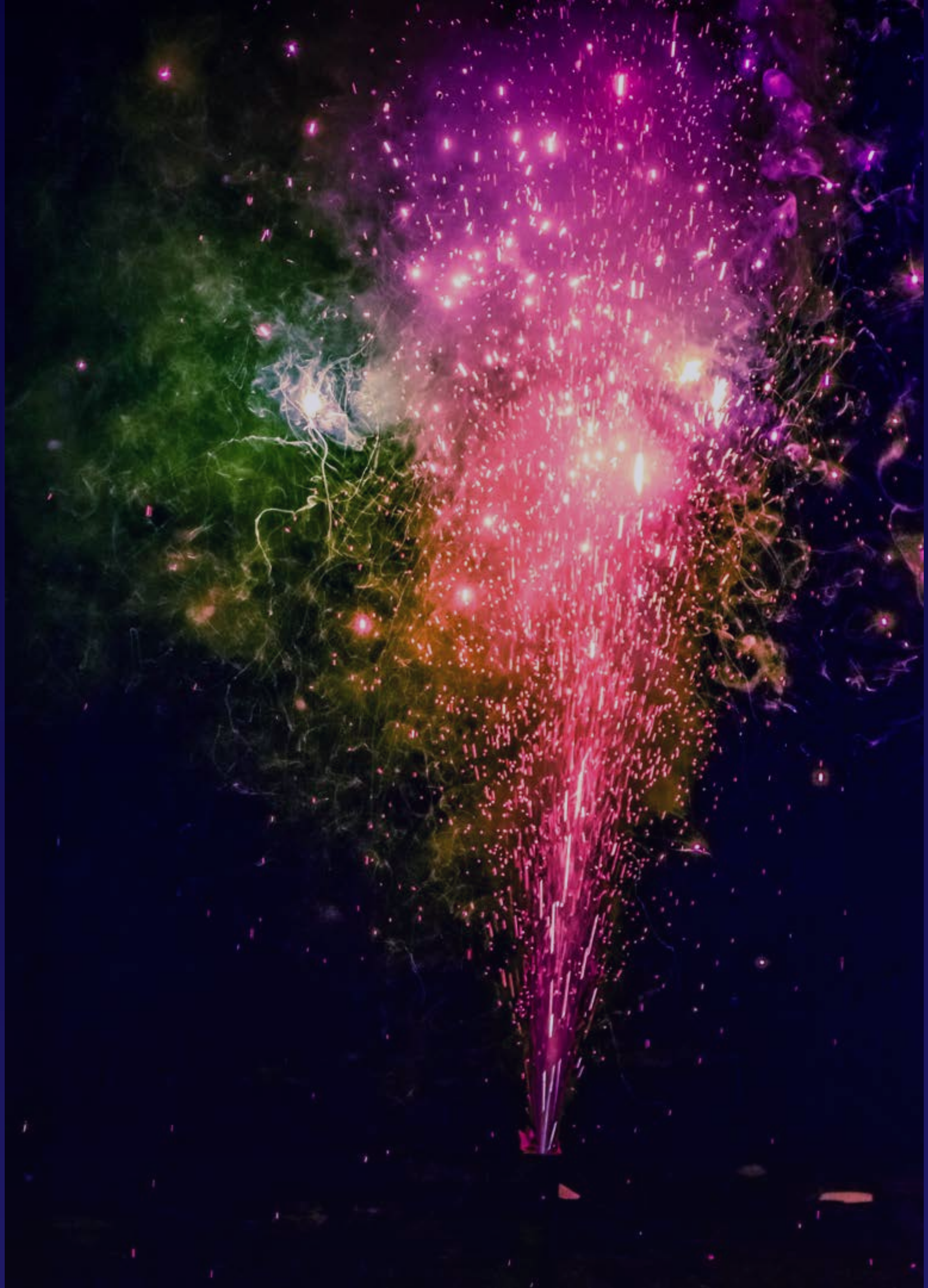
“Scala combines object-oriented and functional programming in one concise, high-level language. Scala's static types help avoid bugs in complex applications, and its JVM and JavaScript runtimes let you build high-performance systems with easy access to huge ecosystems of libraries.”

https://www.scala-lang.org/

Apache Spark Overview

This activity executes a Spark program on either your own or on-demand HDInsight cluster.

Spark jobs are more extensible, allowing you to provide multiple files such as Python scripts and JAR packages.



Notebooks Overview

"Jupyter is a free, open-source, interactive web tool known as a computational notebook, which researchers can use to combine software code, computational output, explanatory text and multimedia resources in a single document."

https://www.nature.com/articles/d41586-018-07196-1

Allows you to use Python and Scala code in Azure Databricks and Azure Machine Learning for data transformations.

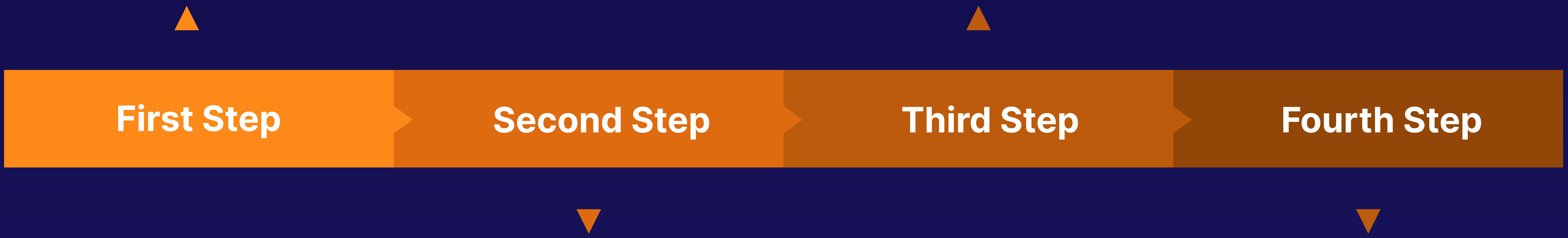
Unit Tests vs. Functional Tests

Write Tests

Using Visual Studio or another IDE, write the unit and functional tests.

Count Activities

Verify how many activities were executed and what status they ended with.



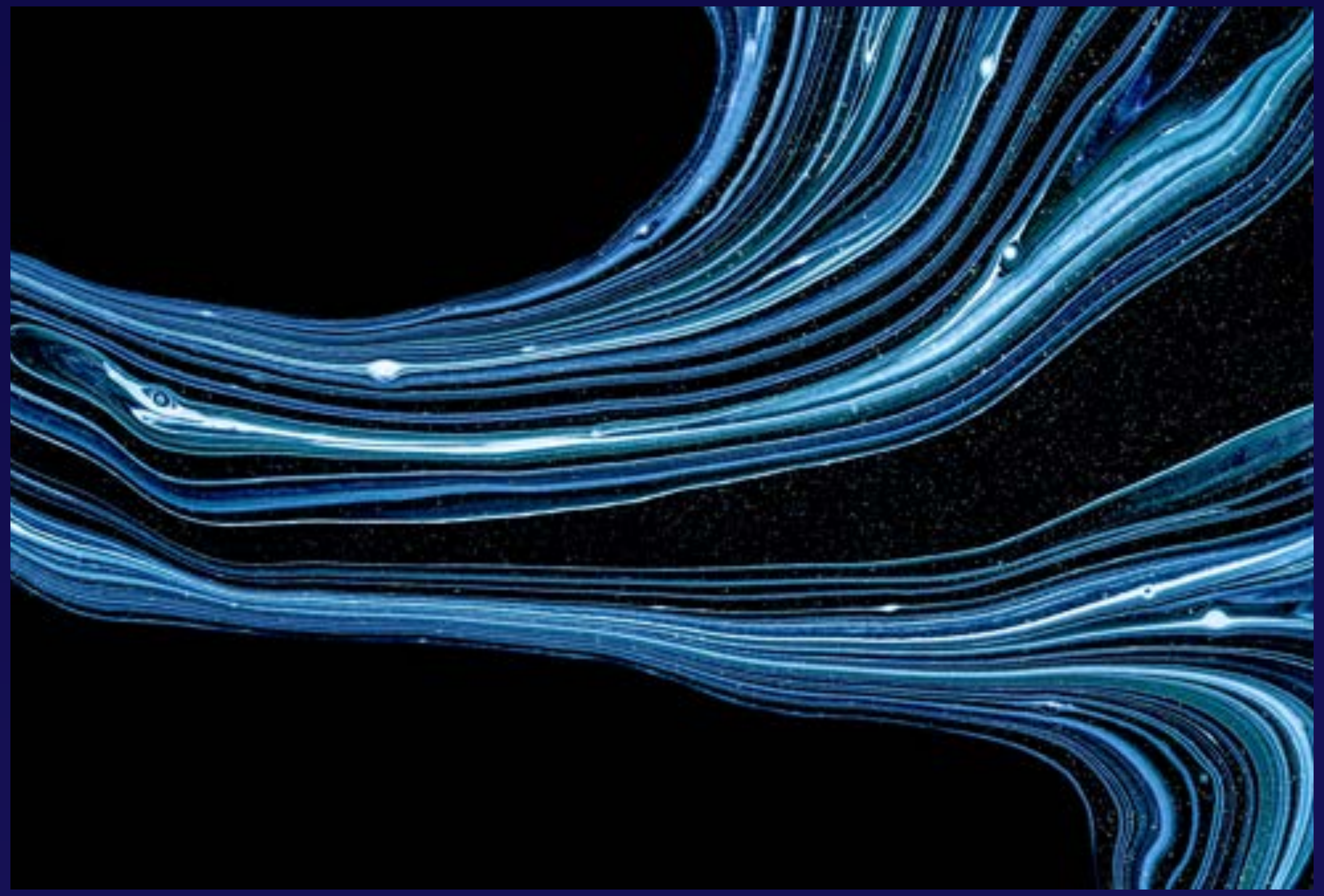
Publish Pipeline

Before tests can be run, the pipeline must be published.

Check Row Counts

Verify the output of activities by counting the rows copied or transformed.

Available Tools



Data Quality Services (DQS)

A component of SQL Server that allows for computer-assisted data cleansing.

Clean Missing Data Module

When working in Azure Machine Learning, this module allows you to replace, remove, and even infer values.

Mapping Data Flows

As part of Azure Data Factory, these activities allow you to include data cleansing as part of your pipeline.

Conditional Split

Routes data rows to particular streams based on specified conditions.

Similar to a CASE statement in traditional programming.

SHREDDING JSON

Code Example

```
SELECT * INTO JsonStudents
FROM OPENJSON(@json, '$.students.azure')
WITH  (
        Id      int             '$.id',
        Name    varchar(60)     '$.name',
        Surname varchar(60)      '$.surname',
        Azure   nvarchar(max)   '$' AS JSON
    );
```

Handling Errors

Continue on Error

Your first, and usually best, option.

Transaction Commit

Choose whether to write data in a single transaction or in batches.

Output Rejected Data

Log the error rows to a CSV in Azure Storage, including the SQL operation and error information.

Success on Error

Mark the data flow as successful even if errors occur.

Exploratory Data Analysis

"Exploratory Data Analysis (EDA) refers to the critical process of performing initial investigations on data so as to discover patterns, to spot anomalies, to test hypothesis and to check assumptions with the help of summary statistics and graphical representations."

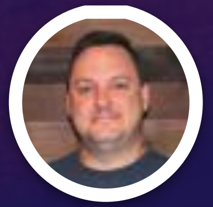
Making sense of the data you have before going too deep with it.

https://towardsdatascience.com/exploratory-data-analysis-8fc1cb20fd15

A CLOUD GURU

Section 5 Review:

Batch Processing Solutions



Brian Roehm

TRAINING ARCHITECT

A CLOUD GURU

Where Is Batch Used?

Banking
Retail
Hospitals
Marketing

Challenges to Consider:

Data Format
Encoding
Dealing with Windows and Missed Runs

Full Data Loading

Dump the Entire Dataset

Completely Replace

No Additional Requirements

vs.

Incremental Data Loading

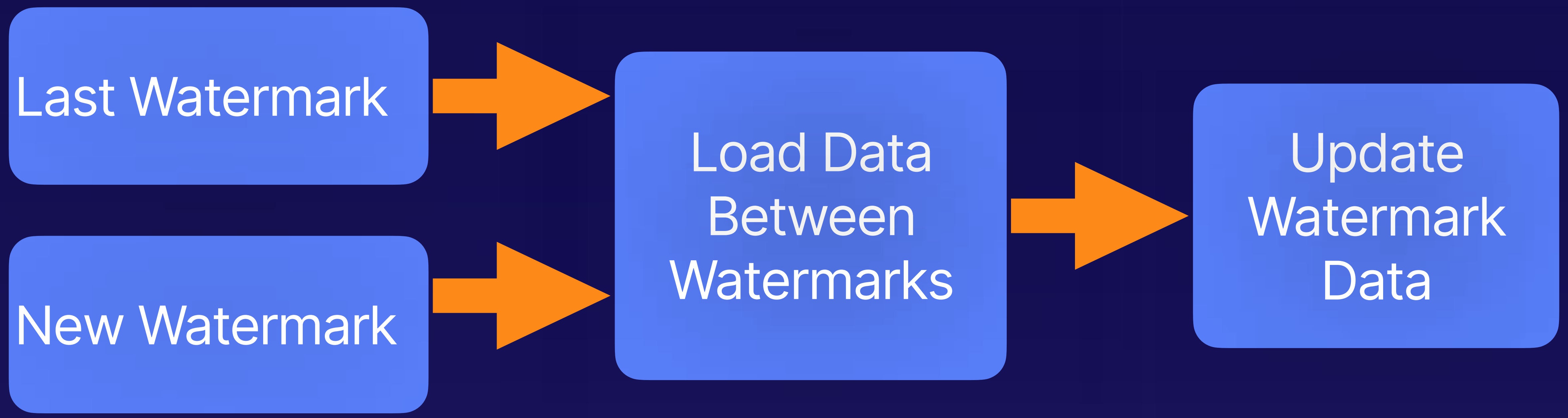
Don’t Dump Anything

Load the Difference Only

1st Tip: Don’t Do This.

The Process

1. Create 2 lookup activities.
2. Create a copy activity.
3. Create a stored procedure activity to update the watermark.



1st Tip: Don't Do This.

Mapping Data Flows

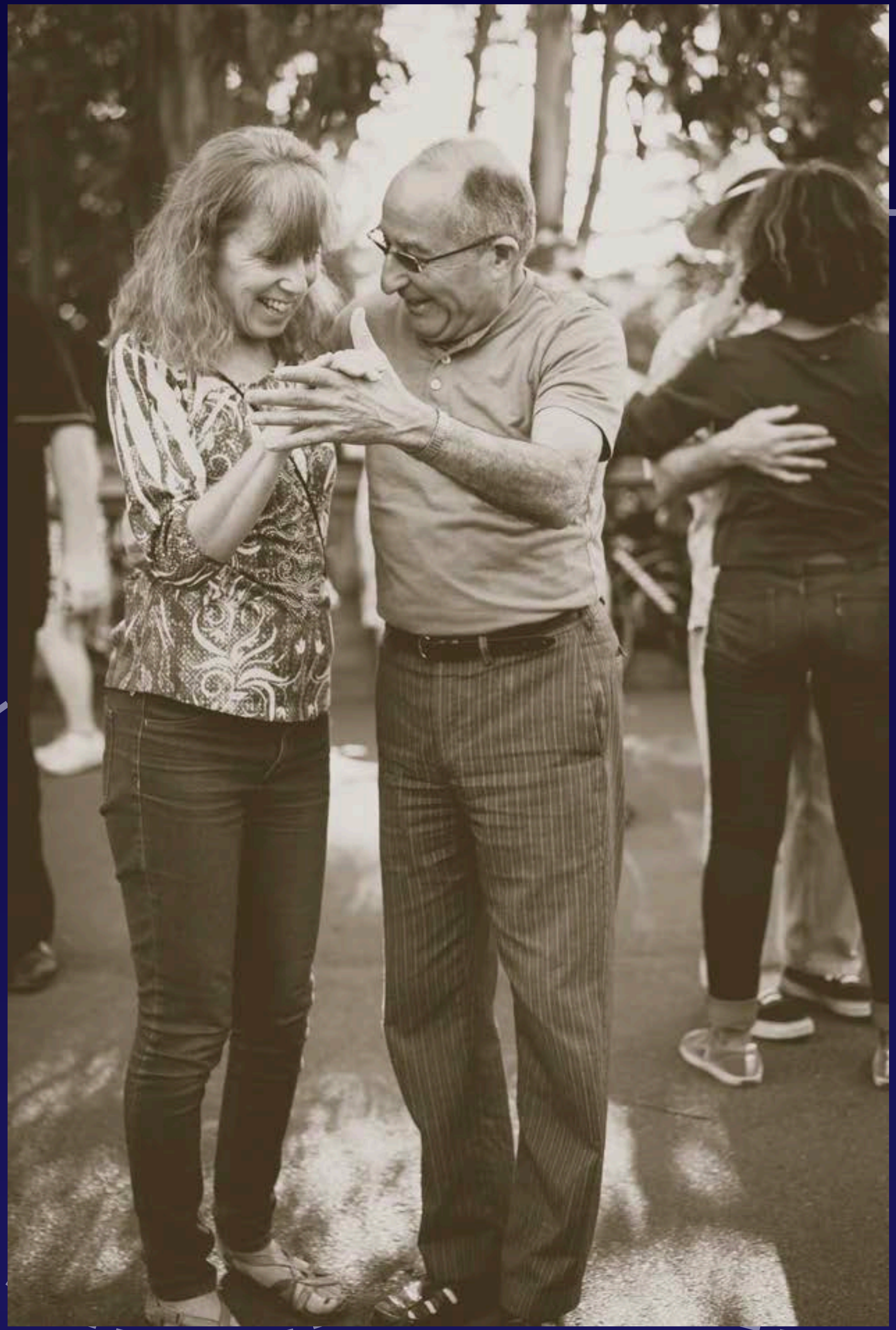
Defined

A visual, no-code solution to developing and implementing transformational logic in Azure Data Factory.

Data flows are created and added into Data Factory pipelines.

The Basic Steps:

1. Create a data flow in Data Factory.
2. Start with the source.
3. Choose your modifier.
4. Choose your destination.



* Make use of scripts as needed.

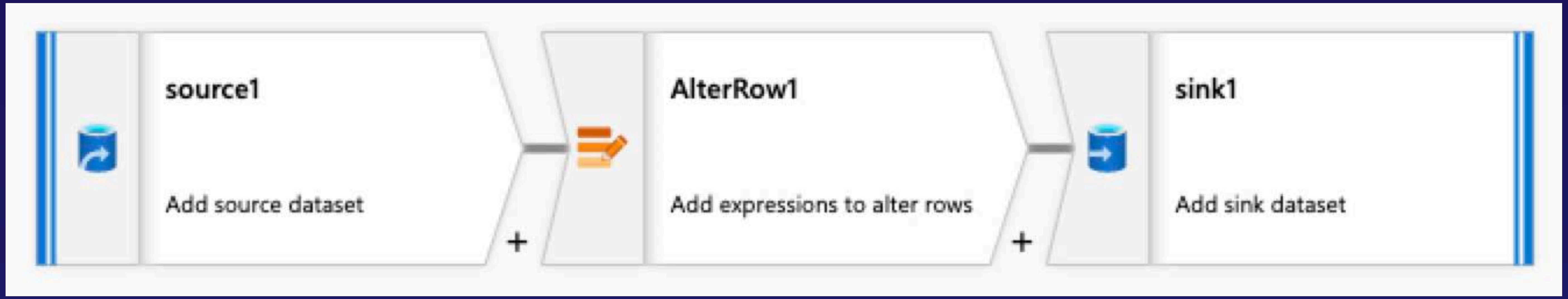
https://docs.microsoft.com/en-us/azure/data-factory/data-flow-script#distinct-row-using-all-columns

Upserting

Definition:

An operation that allows you to either insert rows into a database table if they do not already exist, or update them if they do.

The key is alteration.



source1

Add source dataset

AlterRow1

Add expressions to alter rows

sink1

Add sink dataset

Designing the Exception Handling Strategy

Activity

Success

Failure

Completion

Skipped

Fault Tolerance

What errors can we ignore?

Retry

Defining retry attempts and time limits.

Configuring a Pipeline Execution Trigger

Schedule

Just a simple wall clock

On-Demand Execution

4 Methods for Configuring

.NET | PowerShell | REST | Python

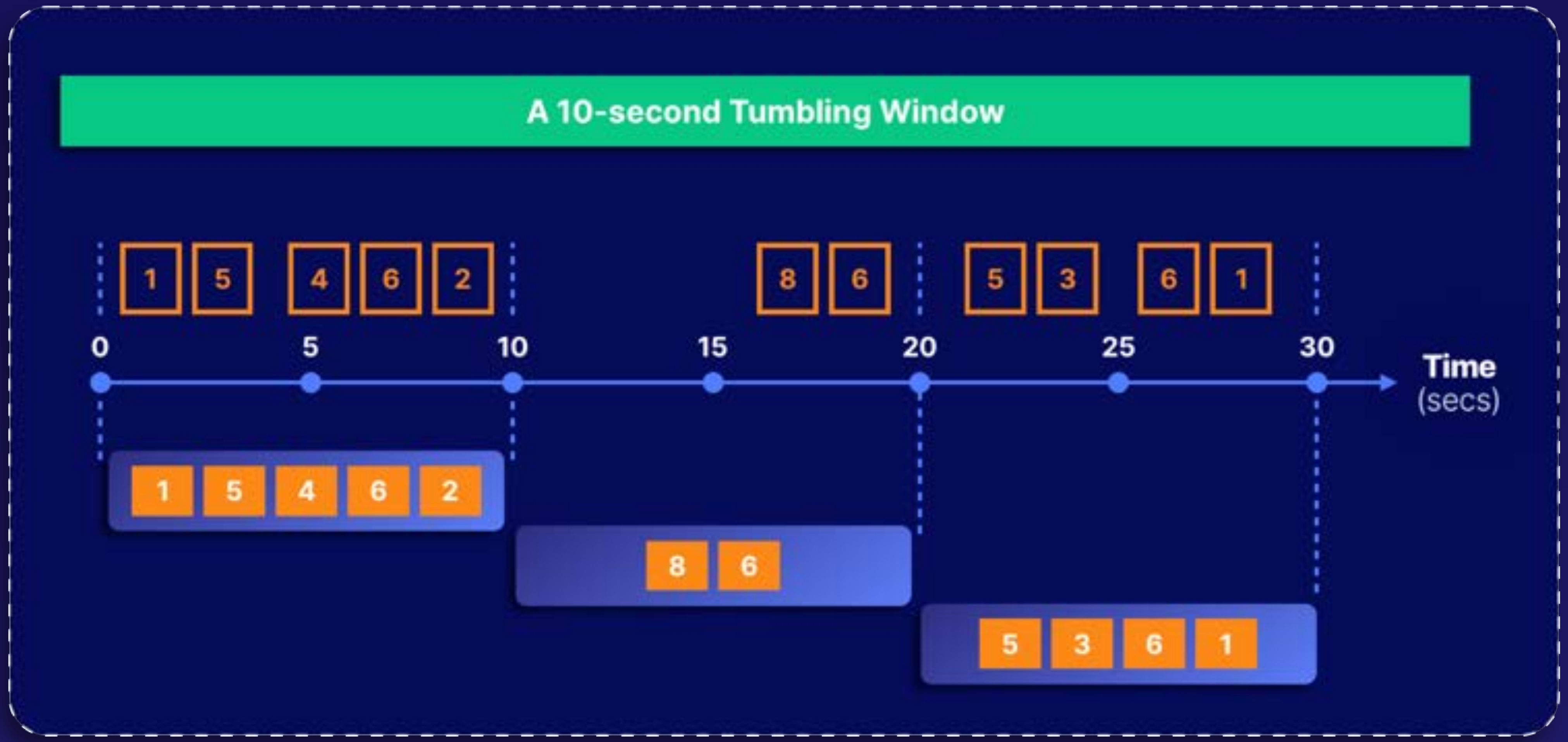
Don’t forget the portal.

Tumbling Window

Fixed

Non-Overlapping

Contiguous Time Intervals



Spark Activity in Data Factory

Execute Spark activities using an HDInsight cluster.

- `sparkJobLinkedService`
  - ONLY Blob Storage and ADLS Gen2
- Script/JAR
  - Just the basics
- `getDebugInfo`
  - None, Always, or Failure (default is None)

```
{
  "name": "Spark Activity",
  "description": "Description",
  "type": "HDInsightSpark",
  "linkedServiceName": {
    "referenceName": "MyHDInsightLinkedService",
    "type": "LinkedServiceReference"
  },
  "typeProperties": {
    "sparkJobLinkedService": {
      "referenceName": "MyAzureStorageLinkedService",
      "type": "LinkedServiceReference"
    },
    "rootPath": "adfspark",
    "entryFilePath": "test.py",
    "sparkConfig": {
      "ConfigItem1": "Value"
    },
    "getDebugInfo": "Failure",
    "arguments": [
      "SampleHadoopJobArgument1"
    ]
  }
}
```

A CLOUD GURU

Section 6 Review:
Stream Processing Solutions

Brian Roehm

TRAINING ARCHITECT

A CLOUD GURU

What Are the Services of Stream Processing?

Hint: There are 3!

1. Azure Stream Analytics
2. Databricks
3. HDInsight

When Would I Use Streaming over Batch?

Streaming is a better fit when you need information now!

It’s great for:

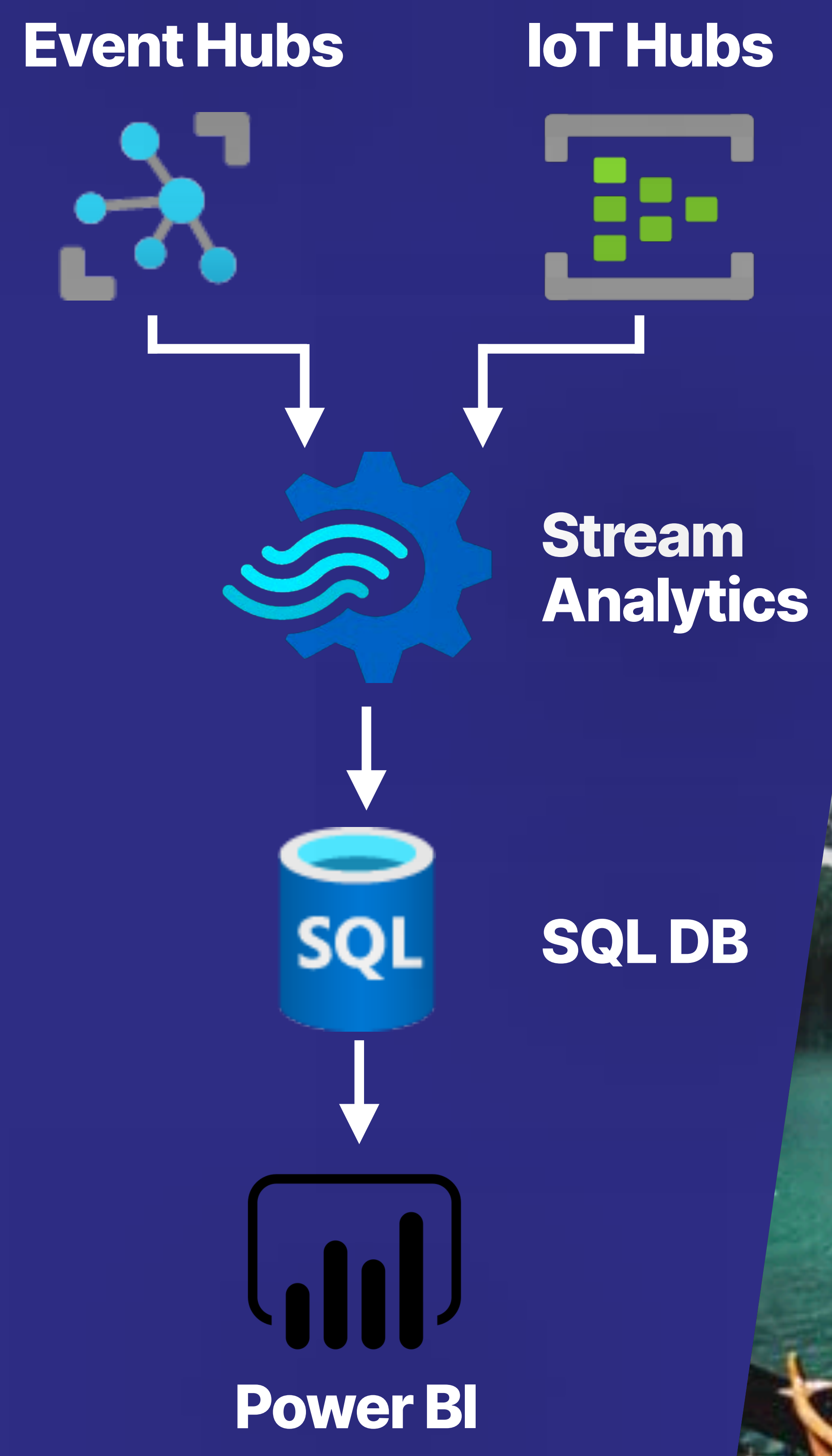
- **Recommendation Engines**
- **Fraud Detection**
- **Some Marketing Applications**
- **Machine Learning**

Don’t forget that streaming generally costs more than batch to run and will allow for less complicated transformation.

IMPORTANT!

There are 5 types of windows:

1. Tumbling
2. Hopping
3. Sliding
4. Session
5. Snapshot



Do You Remember:

Watermarks?

2 Formulas for Watermarks

Watermark = largest event - out-of-order tolerance

Watermark = current estimated arrival - late arrival tolerance

How is time kept?

Event and processing time

The concept of tolerance?

Tolerance set too short can cause data loss

Tolerance set too long can cause a broken process

If I Told You I Wanted to Upsert Data via Stream Analytics, What Would You Tell Me?

Hint: There are 3 conditions!

1. You have to upsert via Cosmos DB.

2. Compatibility level 1.2.

3. It will require configuration on the Cosmos DB Side.

Don’t Forget to Set Alerts and Monitor Your Jobs

1. Understand Key Metrics
   Events, utilization, watermarking, and errors.

2. Create Alerts

Partitioning and Repartitioning

Divide data into subsets based on a partition key.

Why We Partition

Subsets make searching faster.

Partition Keys

Static

High Cardinality (Big Range)

Basics of Repartitioning

For scenarios that aren't fully parallelized

Can process partitions independently

Oh NO! Your Stream Analytics Job Crashed. Should You Panic?

No! Because:

1. Job Is Started

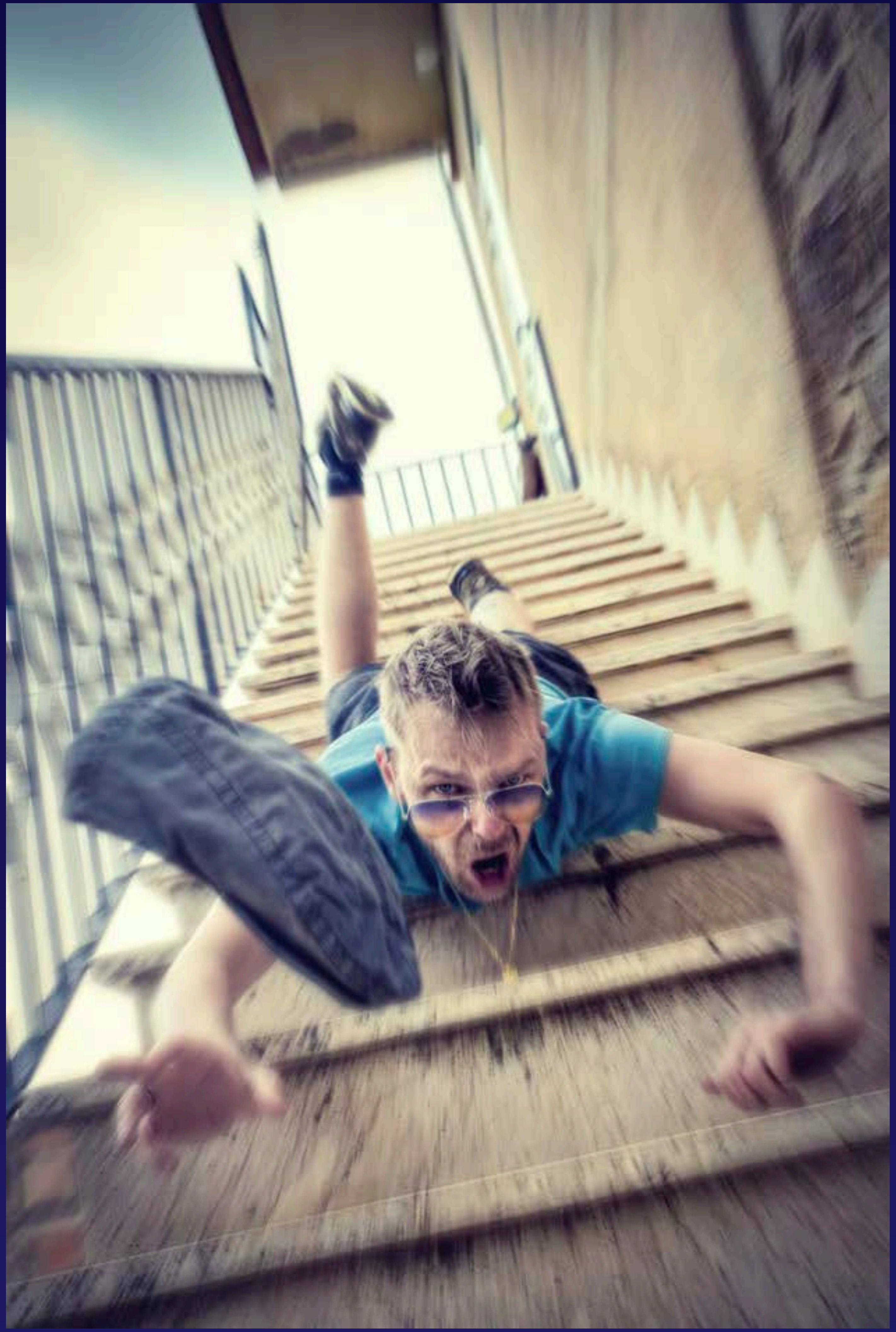
   The work is broken up with worker nodes.

2. Something Bad Happens

   Failure occurs in a node.

3. Automatic Recovery

   Restoration occurs from the last available checkpoint.



Azure Stream Analytics Output Error Policy

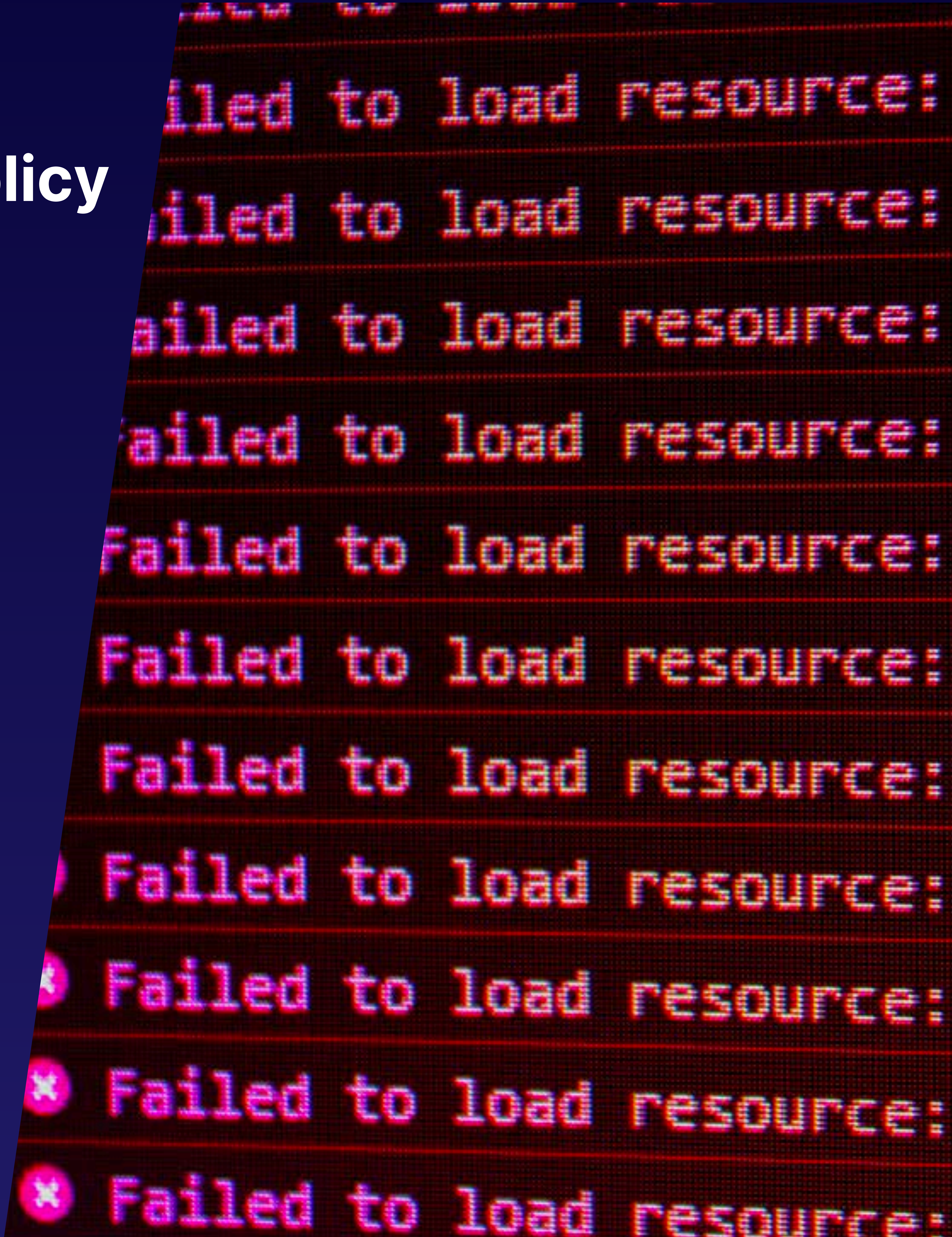
Drop

Drop output events that result in data conversion error.

Retry

Retry the event until the write succeeds

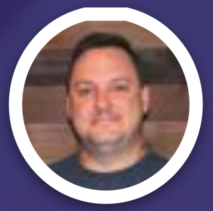
. . . it could LITERALLY take forever.



```
iled to load resource:
iled to load resource:
ailed to load resource:
ailed to load resource:
Failed to load resource:
Failed to load resource:
Failed to load resource:
Failed to load resource:
Failed to load resource:
Failed to load resource:
Failed to load resource:
```

A CLOUD GURU

Section 7 Review:
Data Serving Layer



Brian Roehm
TRAINING ARCHITECT

Star Schema

1 Fact table

A central table full of countable items

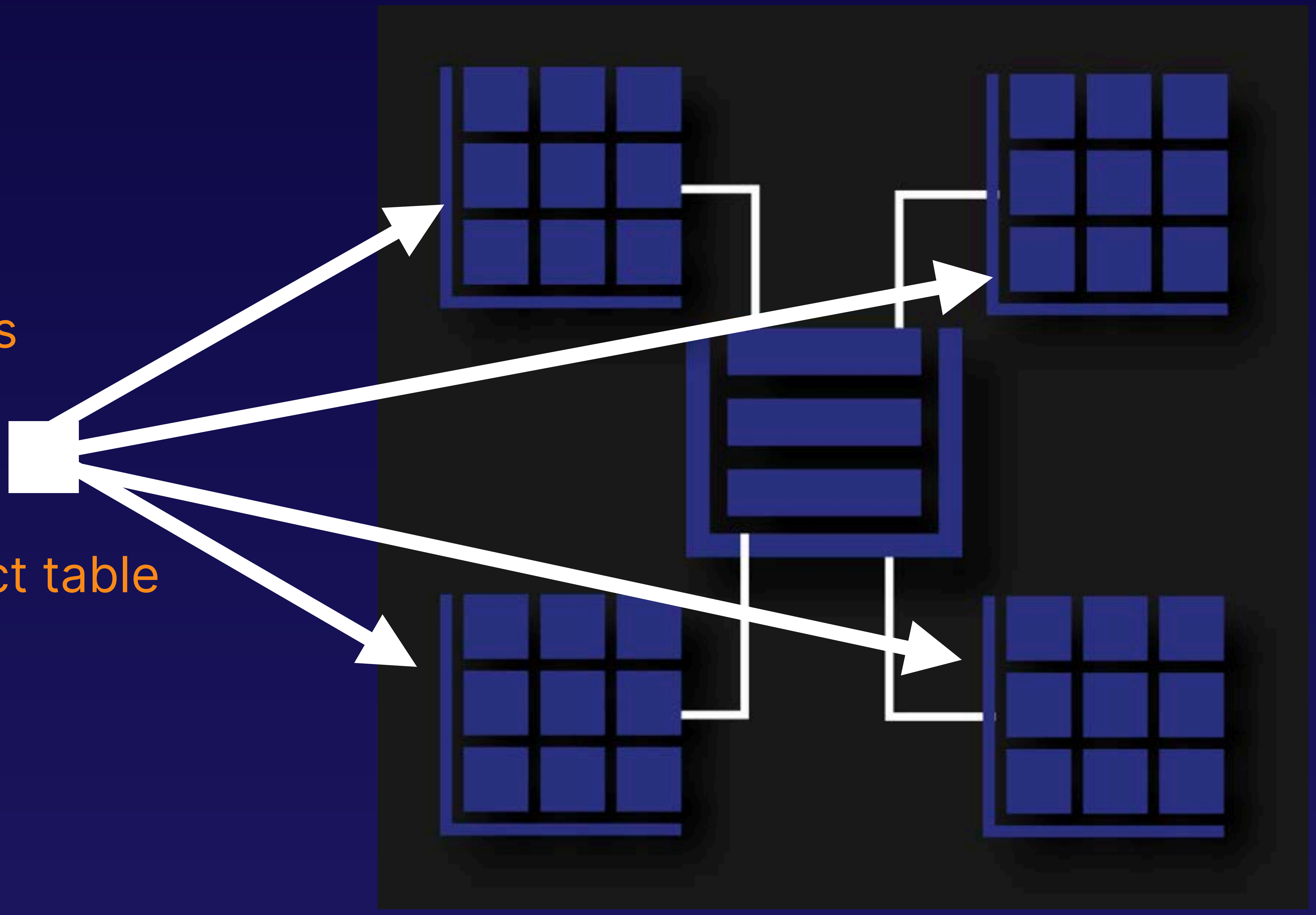
1 dimension table level

All of the other tables tying into the fact table

Not normalized

Think, copies of data.

Easy to query for simple queries



Snowflake Schema

1 Multiple dimension table levels

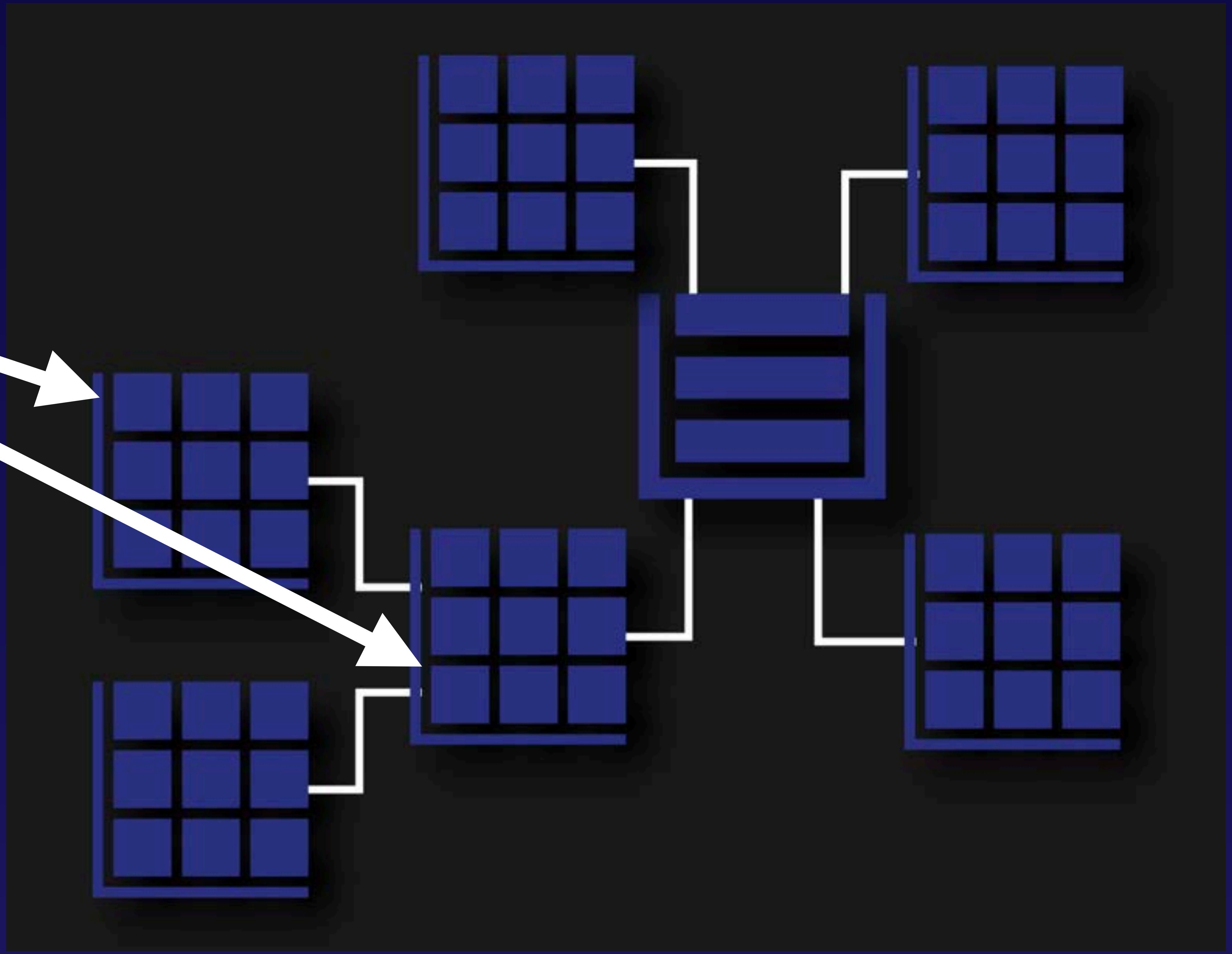
High cardinality

Very little repetition

Normalized

Better for complex queries

Less storage space



Fact Tables and Fact Table Grains

1 Fact Tables

Typically numeric data

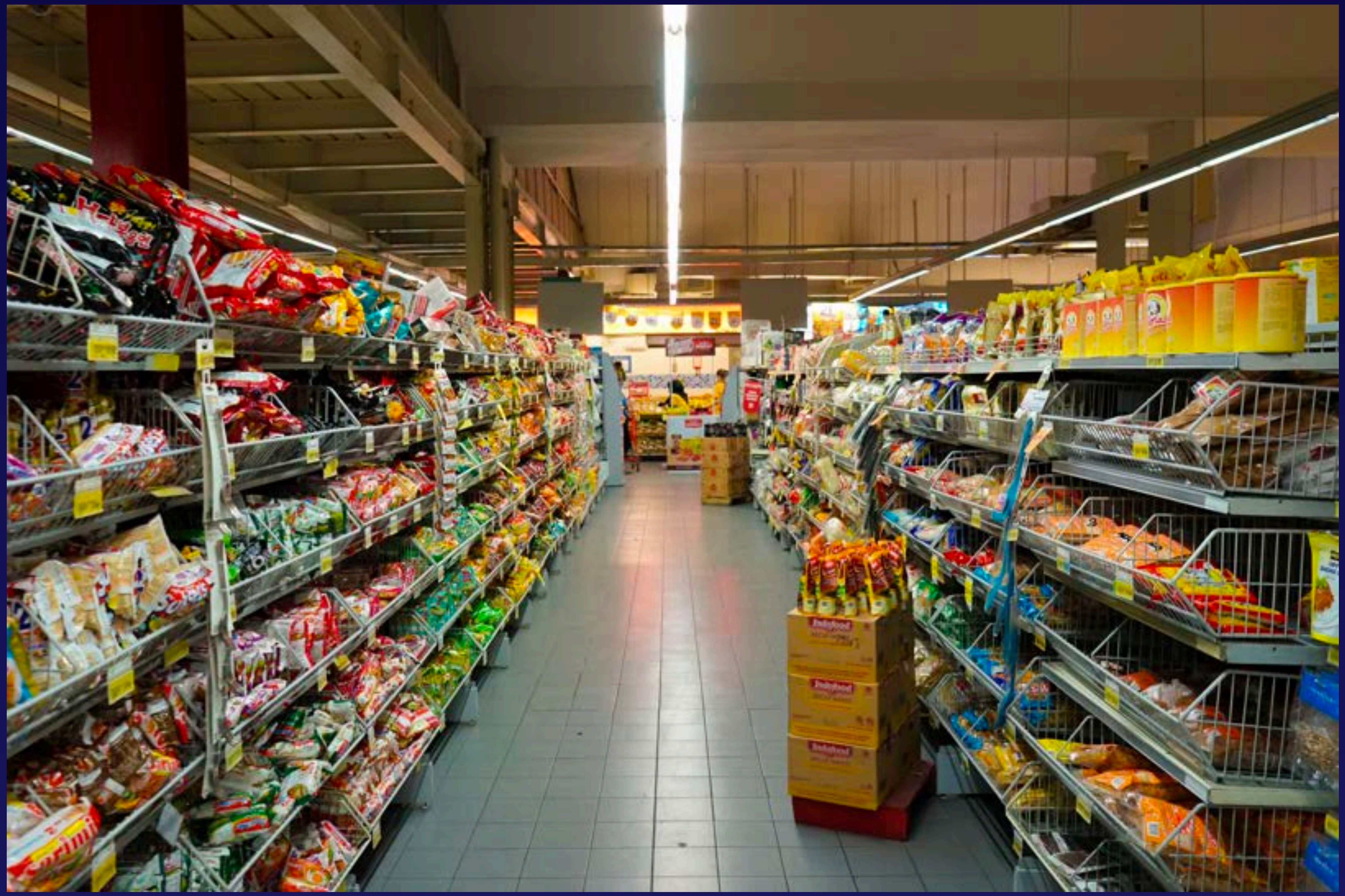
Store: profits, product sales, registers

Each row represents a single event

Measurement data

2 Fact Table Grain

The grain is the level of detail



Relationships between Facts and Dimensions

1 A fact and dimension table have a relationship.

Primary Key

Unique data column used to define relationships.

Foreign Key

Provides a link between data in 2 tables.

What Is an External Table?

1 External tables

Tables whose data comes from files stored outside of the database.

2 But, why?

When you need to access data without needing to copy the ENTIRE dataset.

Provides fast, ad-hoc data access to data hosted outside the bounds of your database.

What Is a Metastore?

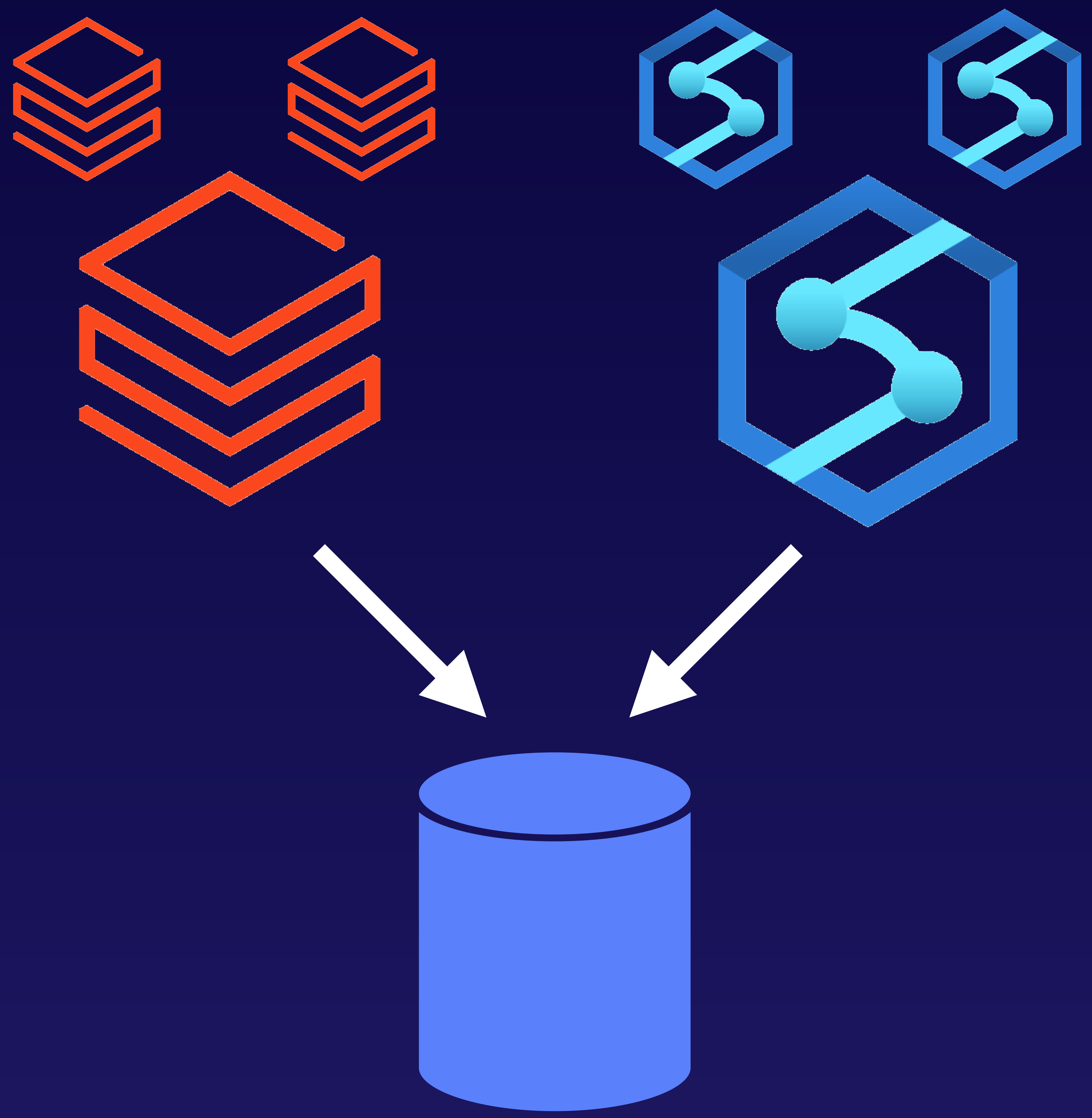
When working in Databricks or Spark, a metastore is created by default. They are only accessible from within a Databricks workspace or Synapse instance (by default).

1 Database that Holds Metadata about Our Data

Paths and formats

2 Can Be in Databricks or Synapse Spark Pools

Metastores can be combined



What, When, Where, and Why?

1. What?

Customer-specified metadata
contentLanguage
contentDisposition

contentType
contentEncoding
cacheControl

2. When and Where?

Data Factory or Synapse pipeline copy activity

3. Why?

Continuity

A CLOUD GURU

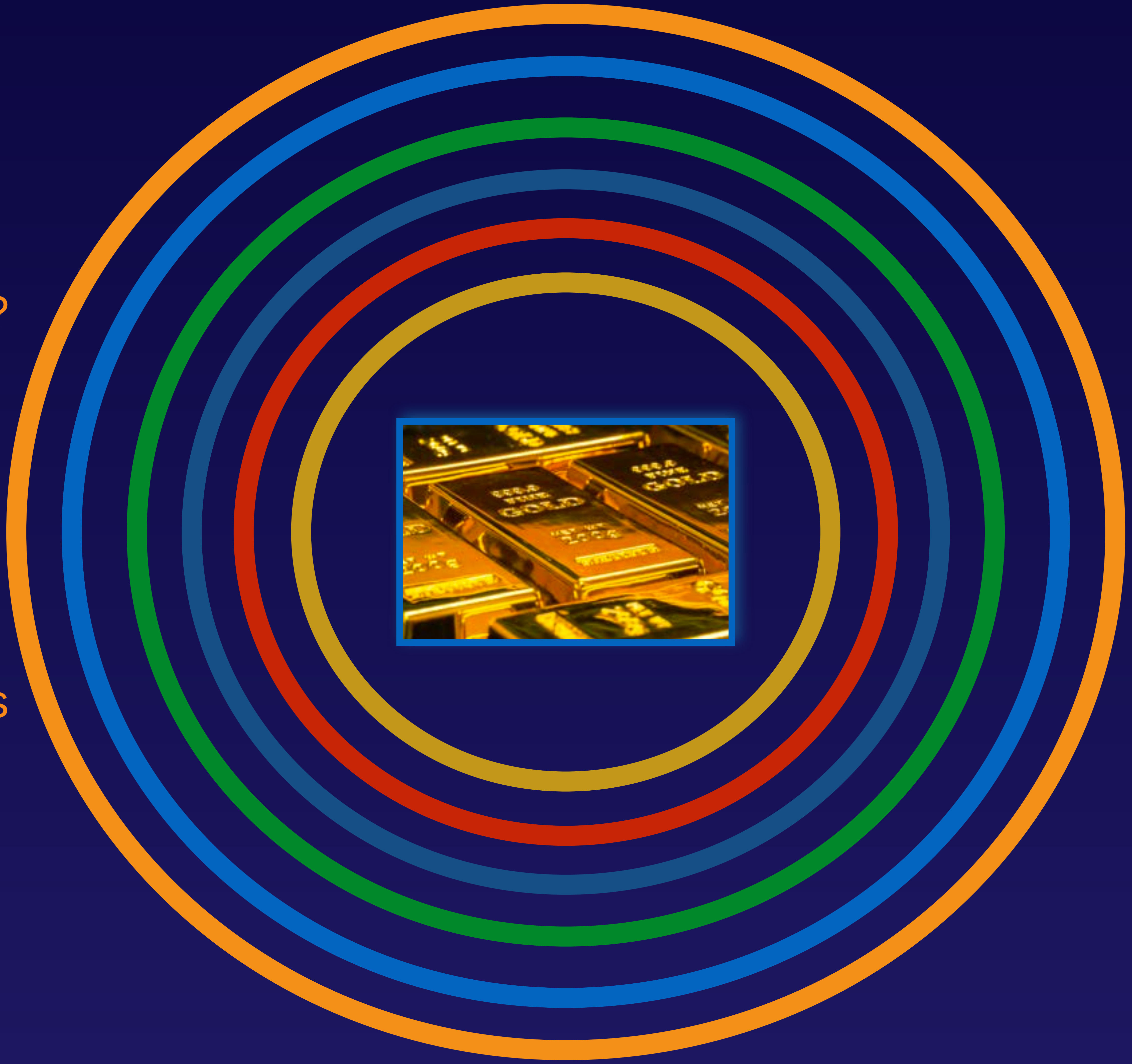
Section 8 Review: Configuring Security and Compliance

Brian Roehm

TRAINING ARCHITECT

Security Layers

1	Physical Security	Building and computing hardware
2	Identify and Access	Who is it and should they be here?
3	Perimeter	Firewalls
4	Network	Network connectivity
5	Compute	VMs and other compute resources
6	Application	A critical point of entry
7	Data	The treasure



Data Encryption

Data at Rest

- Data encryption when not moving
- Symmetric encryption key
  - Most services have this on by default

Data in Motion

- Data encryption when data is moving
- Transport Layer Security (TLS)
  - Most services have this on by default
  - Recommend at least 1.2

When data movement is imperative

Types of Masks

1. Default

   Full masking

2. Credit card

   All but last 4

3. Email

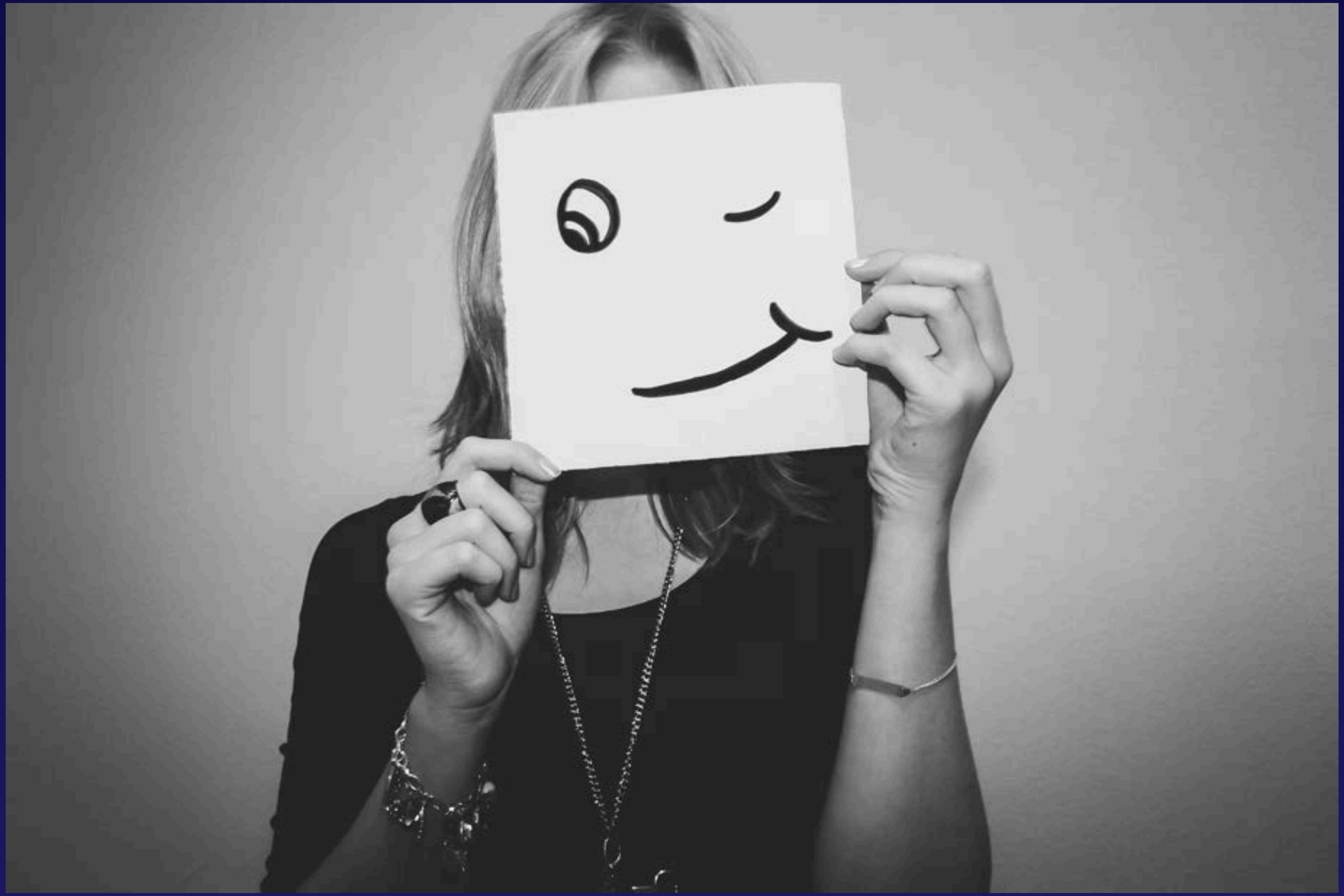
   First letter and .com

4. Random number

   Generate a random number

5. Custom

   Custom padding string



Why We Audit

1. Track database events

   Security concerns

2. Regulatory compliance

   Not the only factor

3. Trends

   Security and operations

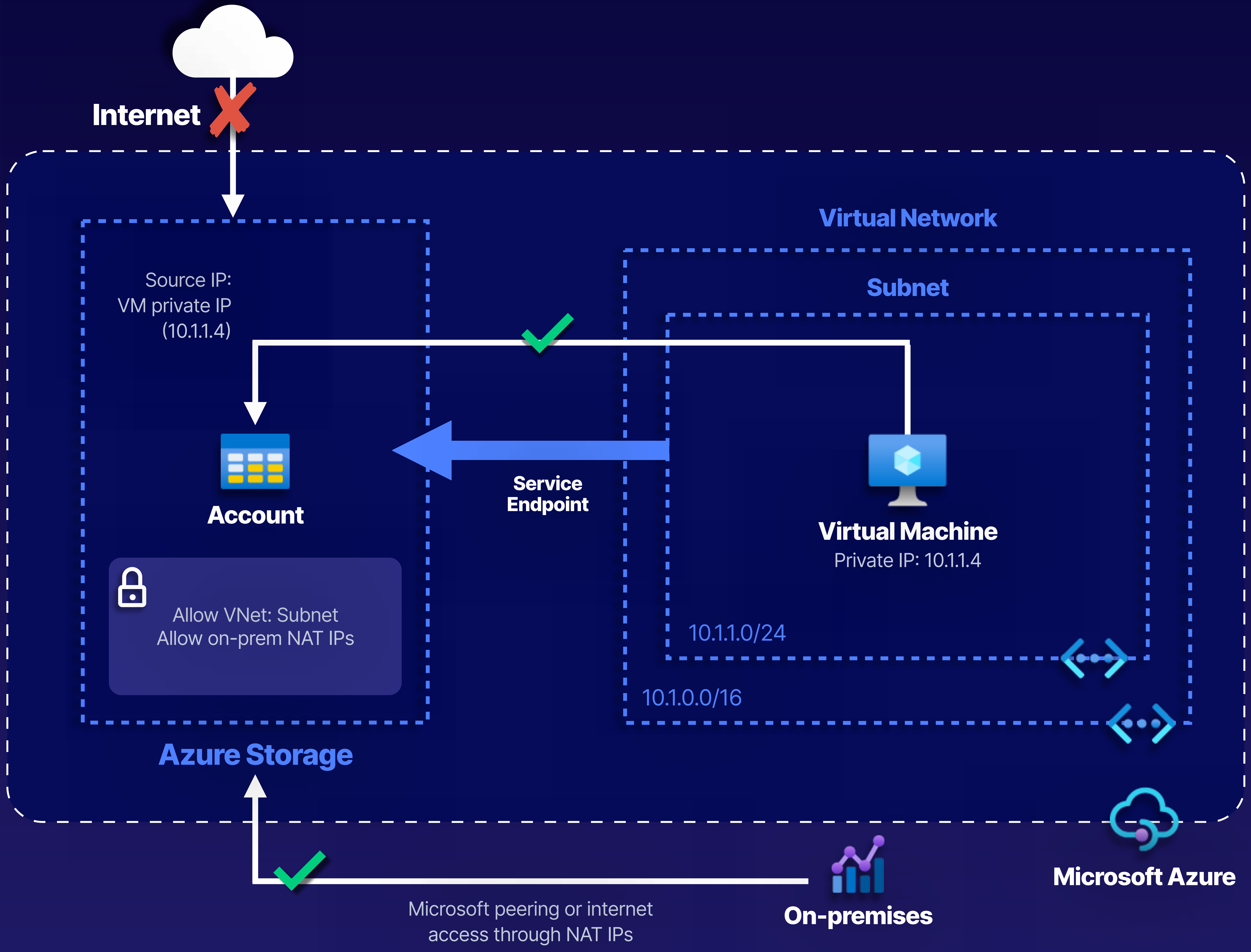
Where We Audit

3. Synapse and SQL DB (Feature)

   Several other places as a practice

Service Endpoints

- Provides secure and direct connectivity to Azure services
- Only for endpoints in Azure virtual networks
- Only for traffic within a virtual network region



Best Practice

1. You need a data retention policy.

   Including lifetime and regulatory requirements

2. Data should move from active to archive to purge.

3. Watch preconfigured backups from services.

4. Plan your purging strategy for off hours.

5. Assess and run cost management analysis.

   Including storage and movement

6. For multicloud and hybrid environments, map storage.

Role-Based Access Control (RBAC)

1 Security principal

To whom does it apply?

2 Role definition

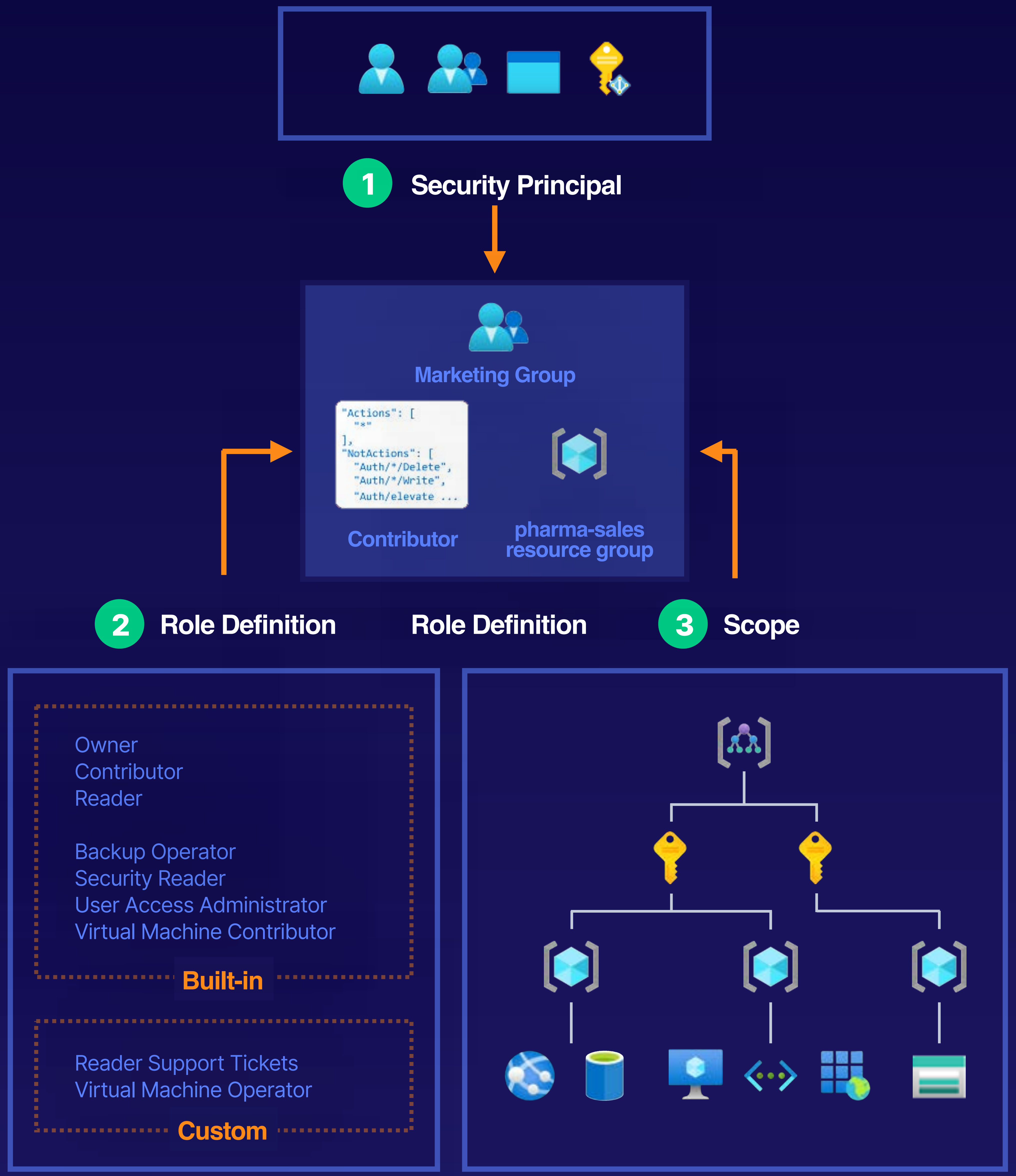
Collection of permissions.

3 Scope

What resources does the role apply to?

4 Role assignment

The marriage of all 3 principals.



1 Security Principal

Marketing Group

```
"Actions": [
  "*"
],
"NotActions": [
  "Auth/*/Delete",
  "Auth/*/Write",
  "Auth/elevate ...
```

Contributor

pharma-sales resource group

2 Role Definition

Role Definition

3 Scope

Owner
Contributor
Reader

Backup Operator
Security Reader
User Access Administrator
Virtual Machine Contributor

Built-in

Reader Support Tickets
Virtual Machine Operator

Custom

Azure Key Vault

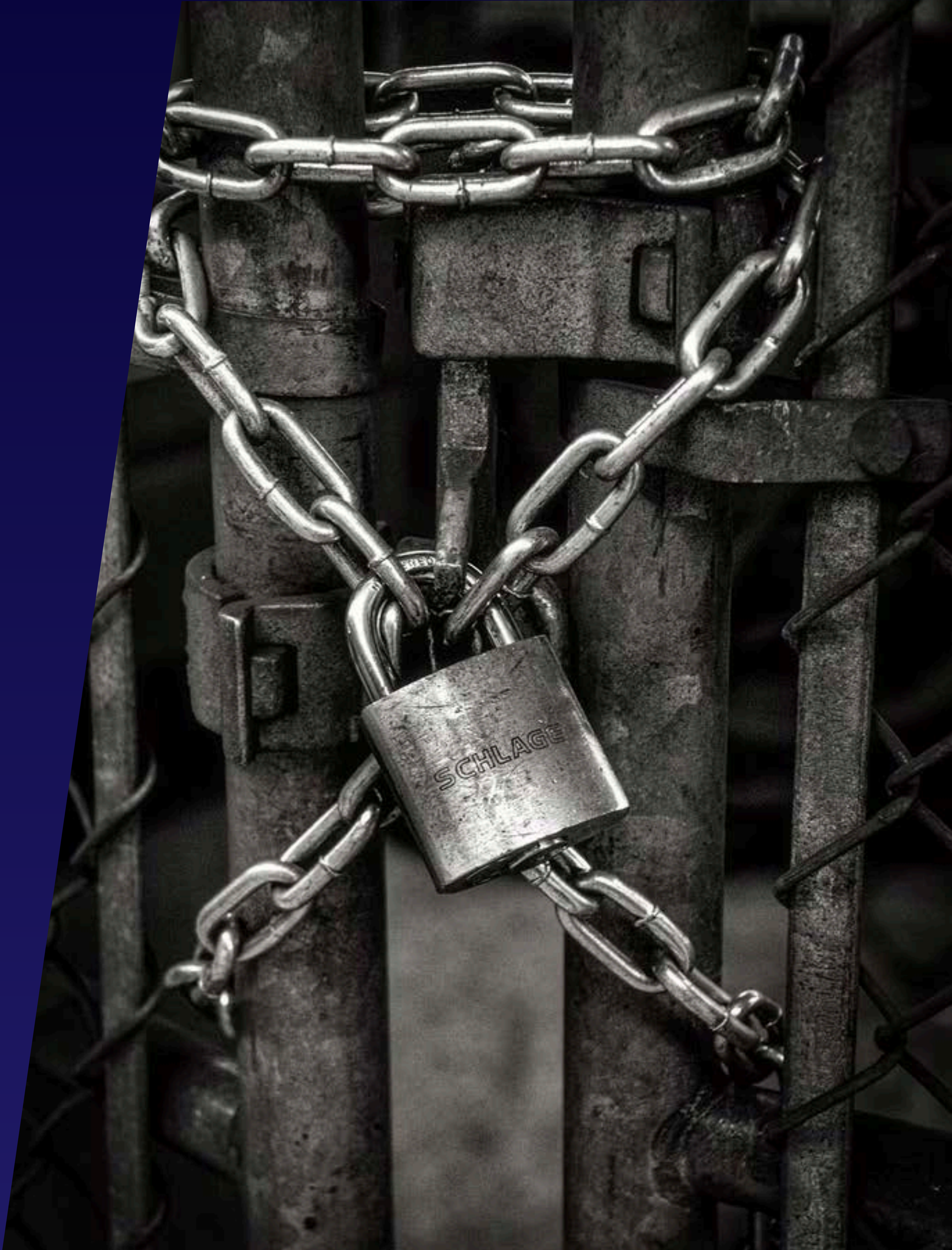
1 **Securely store and access**

Tokens	Certificates
Passwords	API keys

2 **Centralize storage of secrets**

Anything you want to control access to

3 **Monitor access**



2 Options for Authentication in Databricks

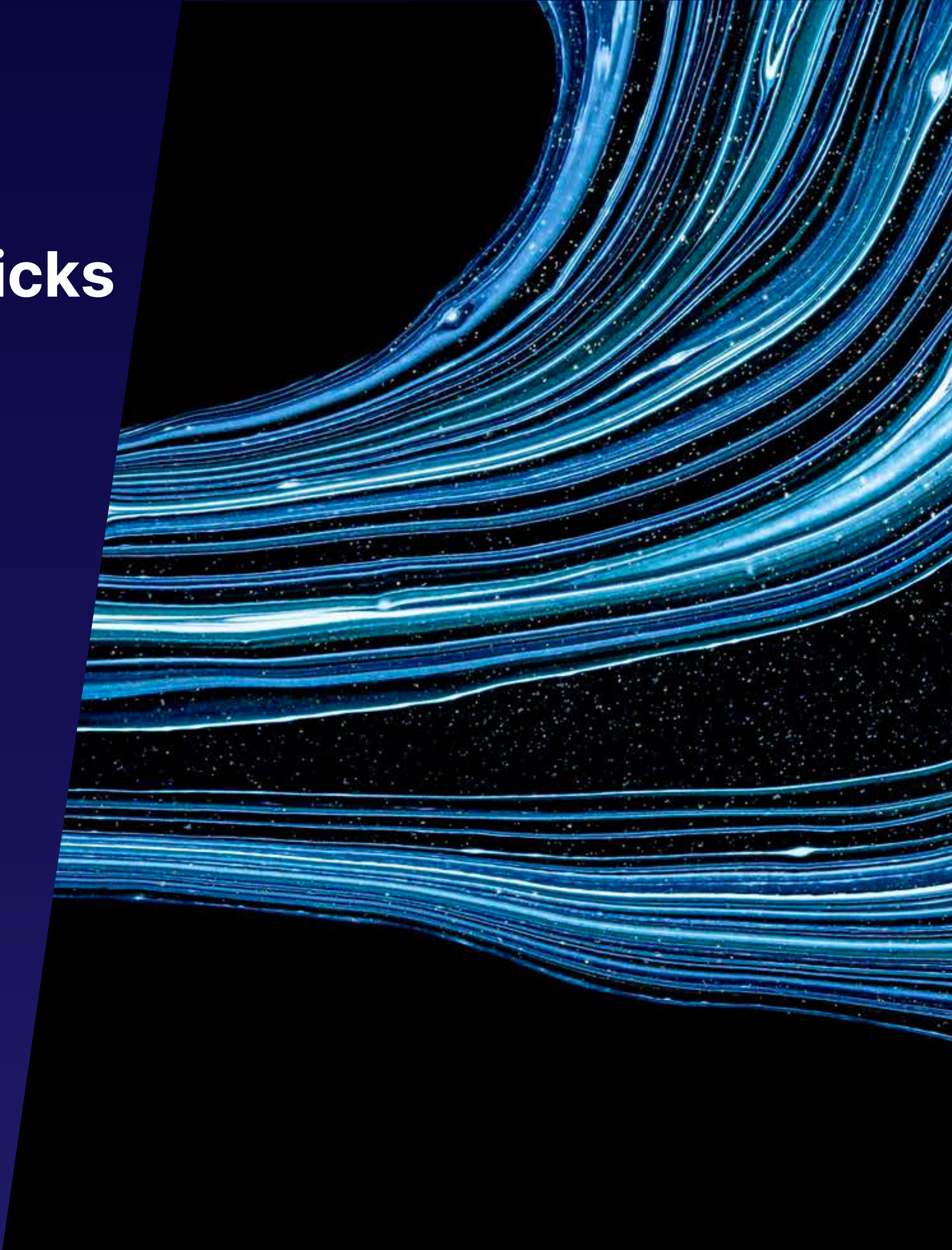
1. Azure Active Directory (Azure AD) tokens

   Define a service principal in Azure AD.

   Get an Azure AD access token.

2. Azure Databricks personal access tokens

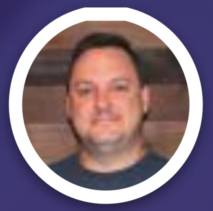
When looking at authentication, tokens should be used in place of passwords.



A CLOUD GURU

Section 9 Review:

Monitor Data Storage and Processing



Brian Roehm

TRAINING ARCHITECT

How Do We Start Collecting Data?

1 Great news! You already are!

2 Azure Monitor generates:

- Activity log
- Platform metrics
- Resource logs
- VM guest metrics and logs

Measuring Performance of Data Movement

1. Estimate > compare > optimize as needed
2. Full utilization allows throughput estimation.
3. Monitor copy data activity to measure performance activity.

Take a Walk in the Synapse Query Store

1 What is the Query Store?

- Provides insight on query plan choice and performance.
- Captures queries, plans, and runtime statistics.

2 Don’t forget to turn it on...

```
ALTER DATABASE <database_name>
SET QUERY_STORE = ON;
```

3 The 3 stores:

- Plan store
- Runtime stats store
- Wait stats store

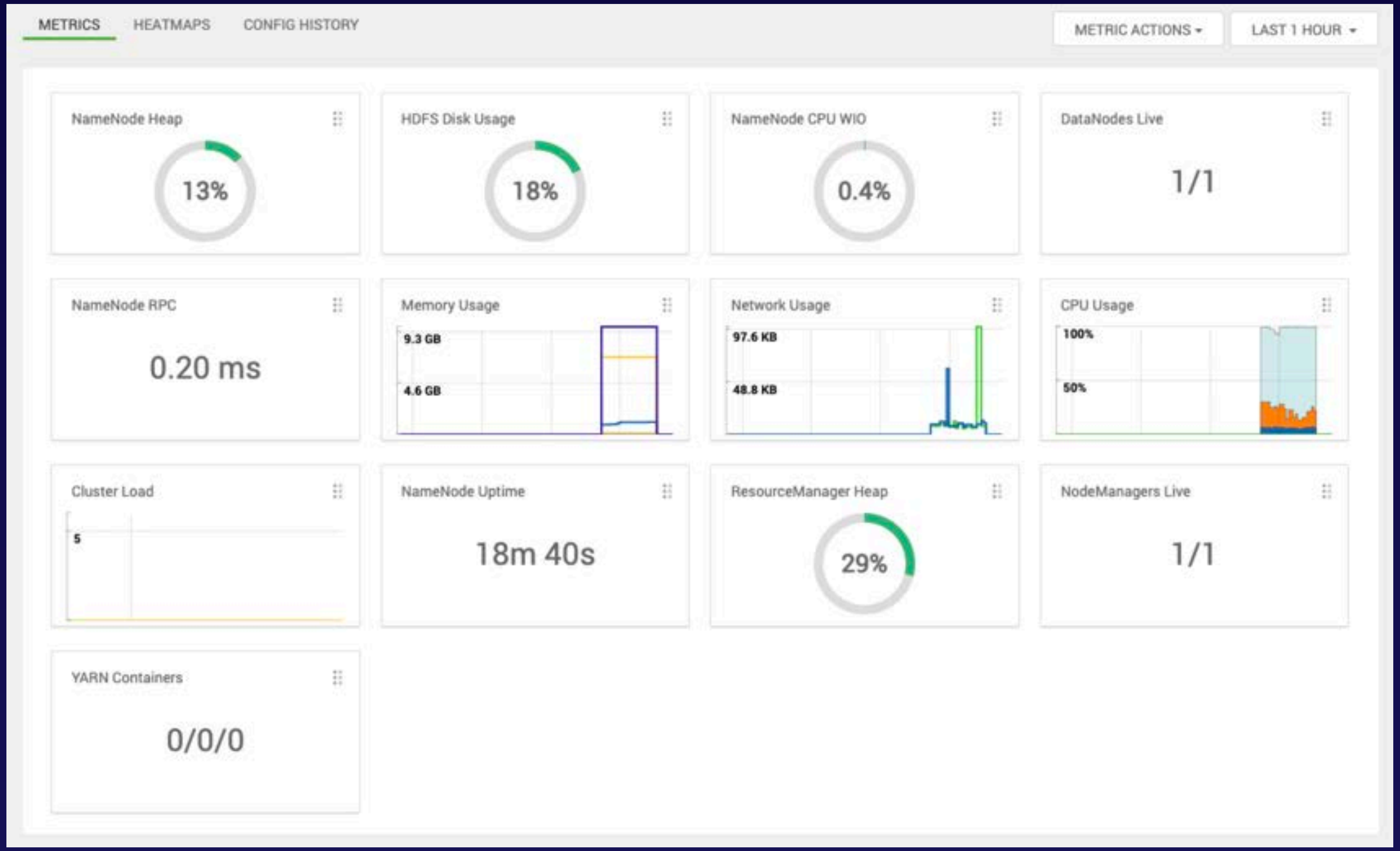
Measuring HDInsight Clusters

1 It all starts with Apache Ambari.

- Manage and monitor HDInsight clusters.
- Create alerts.

2 How do I access Ambari?

https://CLUSTERNAME.azurehdinsight.net



NameNode Heap 13%

HDFS Disk Usage 18%

NameNode CPU WIO 0.4%

DataNodes Live 1/1

NameNode RPC 0.20 ms

Memory Usage

Network Usage

CPU Usage

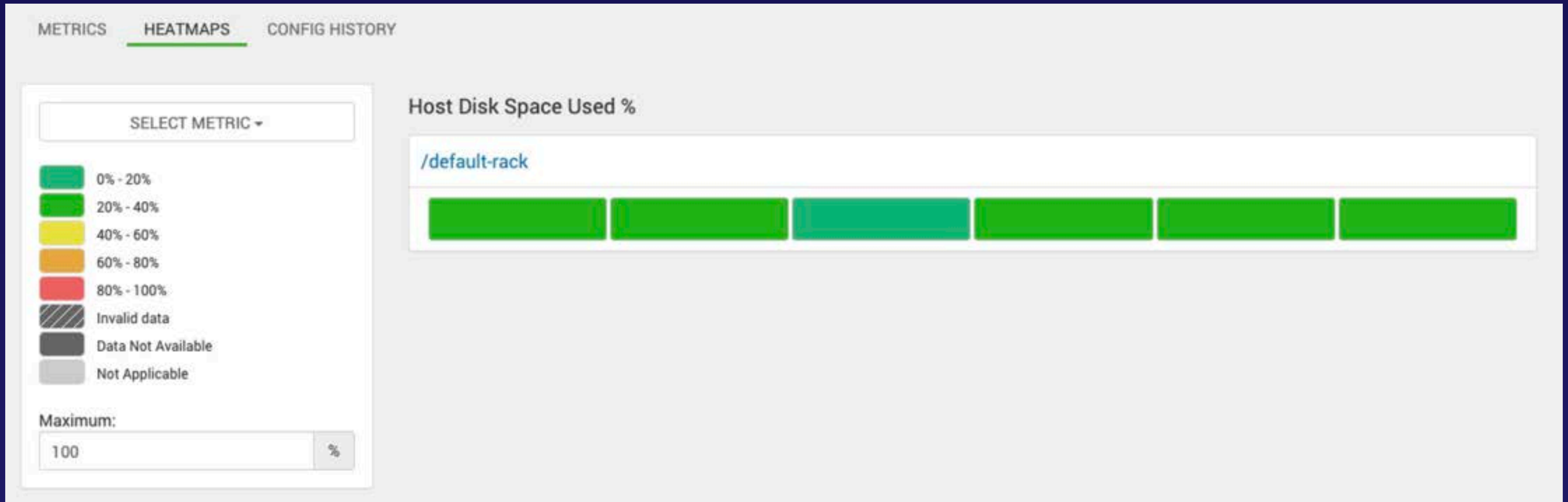
Cluster Load

NameNode Uptime 18m 40s

ResourceManager Heap 29%

NodeManagers Live 1/1

YARN Containers 0/0/0



Host Disk Space Used %

/default-rack

🟩 0% - 20%
🟩 20% - 40%
🟨 40% - 60%
🟧 60% - 80%
🟥 80% - 100%
▨ Invalid data
⬛ Data Not Available
⬜ Not Applicable

Maximum: 100 %

A CLOUD GURU

Section 10 Review:

Optimize and Troubleshoot Data Storage and Processing

Brian Roehm

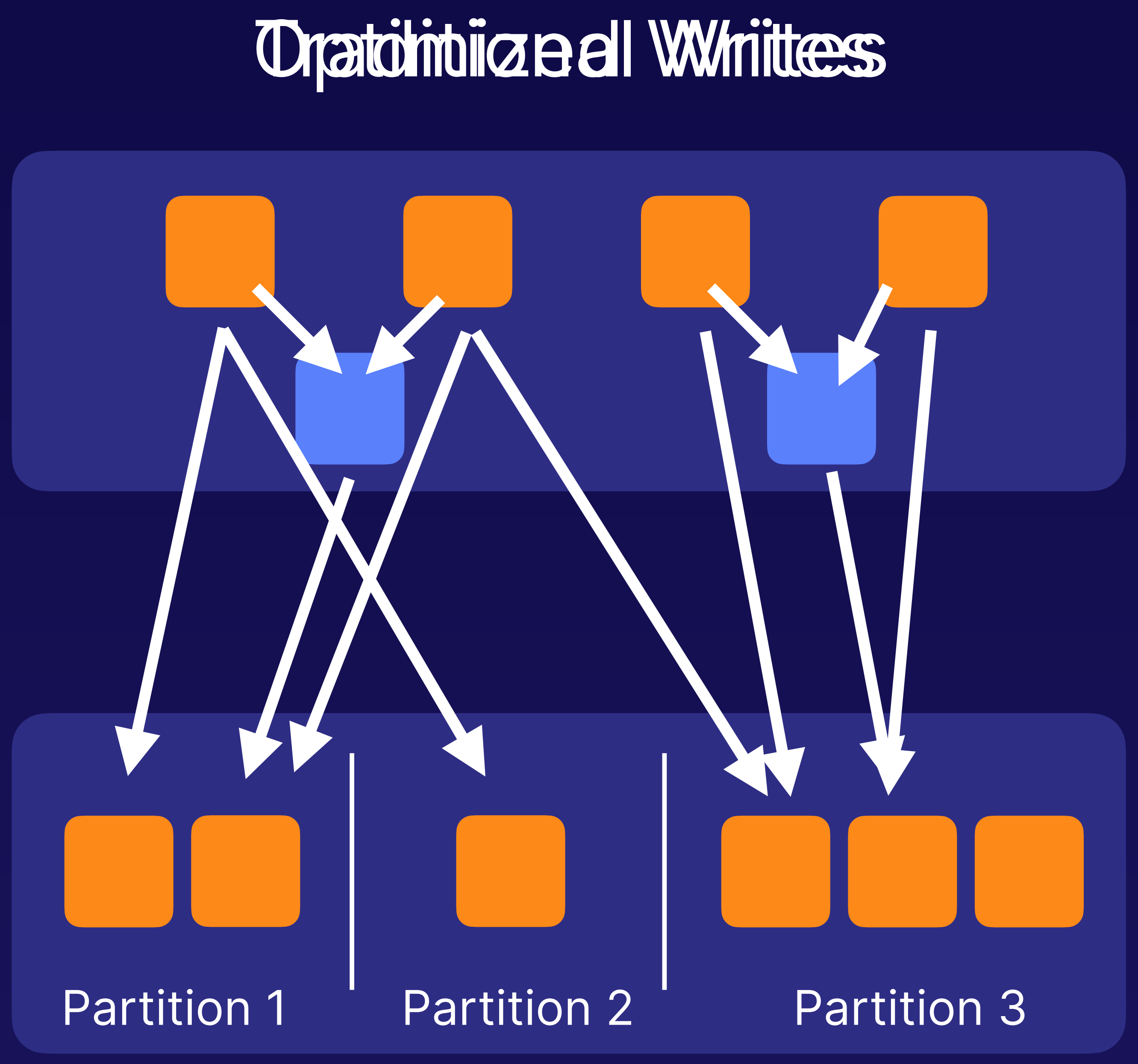
TRAINING ARCHITECT

Auto Optimization Overview

1 This is an Azure Databricks tool.

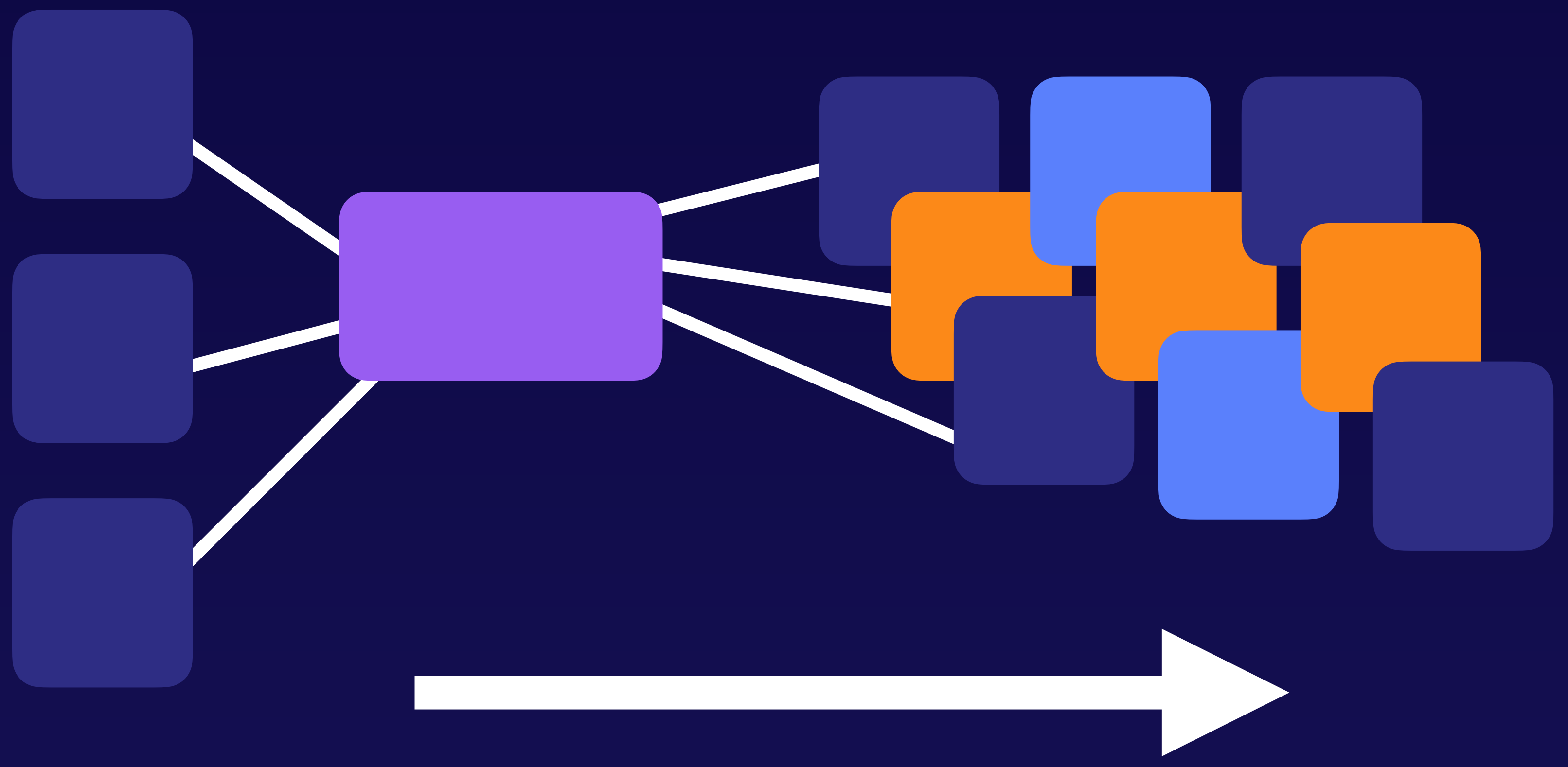
2 Allows you to automatically compact small files.

Individual writes to a Delta table.



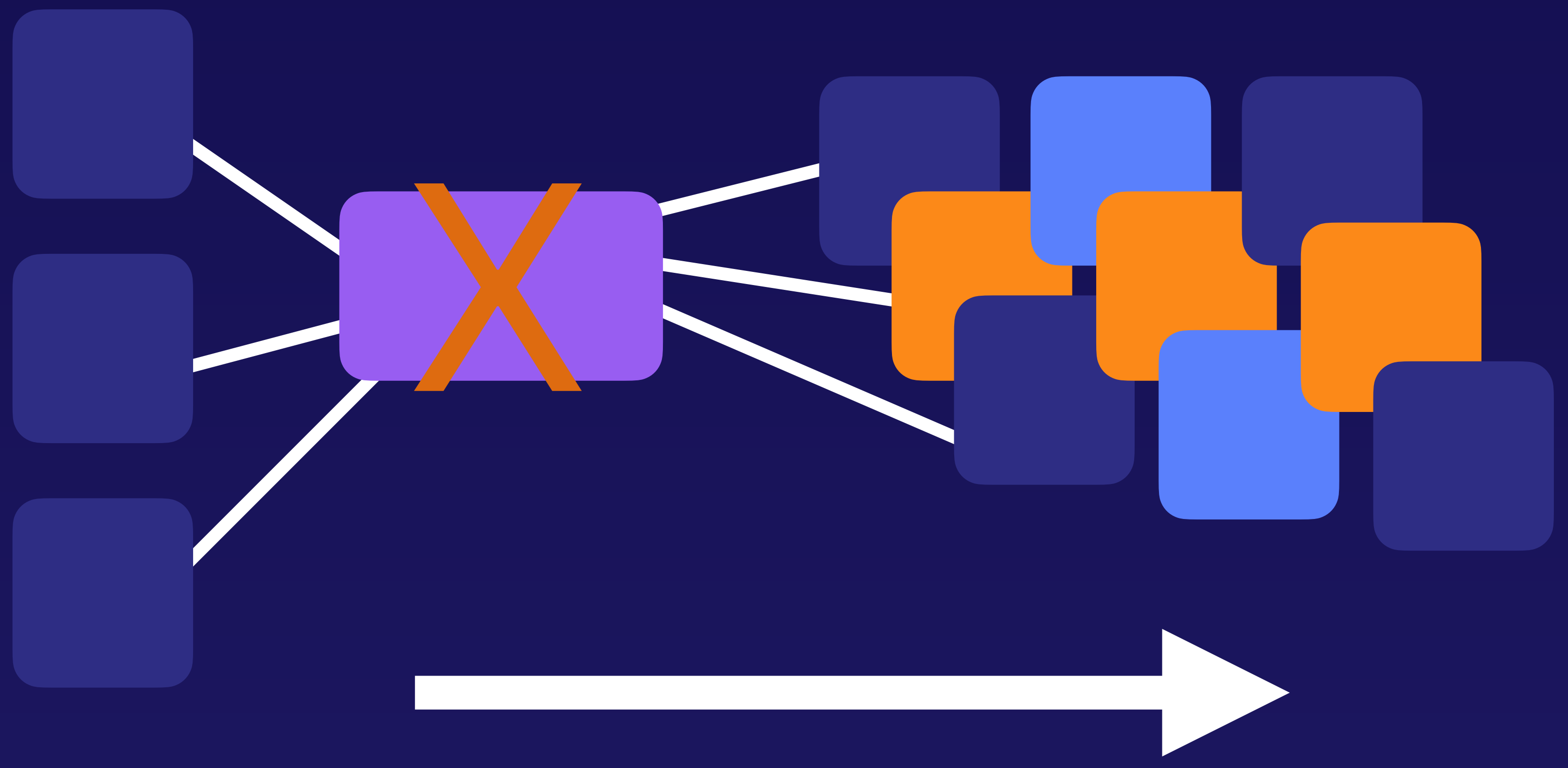
Hash Distribution

1. Hash function used to assign to rows
   - Essentially a fancy math algorithm



Round-Robin Distribution

2. First random, then sequential
   - Quick to load
   - Slow to query



Setting the Shuffle Partition Size

1 The challenge

- Finding the right shuffle partition number

2 Adaptive query execution (AQE) helps to solve this issue

- You set the initial shuffle partition number

```
spark.conf.set("spark.sql.shuffle.partitions",100)
println(df.groupBy("_c0").count().rdd.partitions.length)
```

A CLOUD GURU

Solution 1: Query Plan

1 Don't forget about statistics.

Solution 2: Reducer and Combiner

2 Recursive reducer

The short of it is parallel performance.

```
[SqlUserDefinedReducer(IsRecursive = true)]
```

3 Row-level combiner

The short of it is parallel performance.

```
[SqlUserDefinedCombiner(Mode = CombinerMode.Right)]
public class WatsonDedupCombiner : ICombiner
{
  public override IEnumerable<IRow>
    Combine(IRowset left, IRowset right, IUpdatableRow output)
   {
  //Your combiner code goes here.
  }
}
```

A CLOUD GURU

Optimization Tips for Your Environment

1. Use ARM templates.

   Replication

   Control

2. Don’t put everything in the same subscription.

   Defense in depth

   Compliance

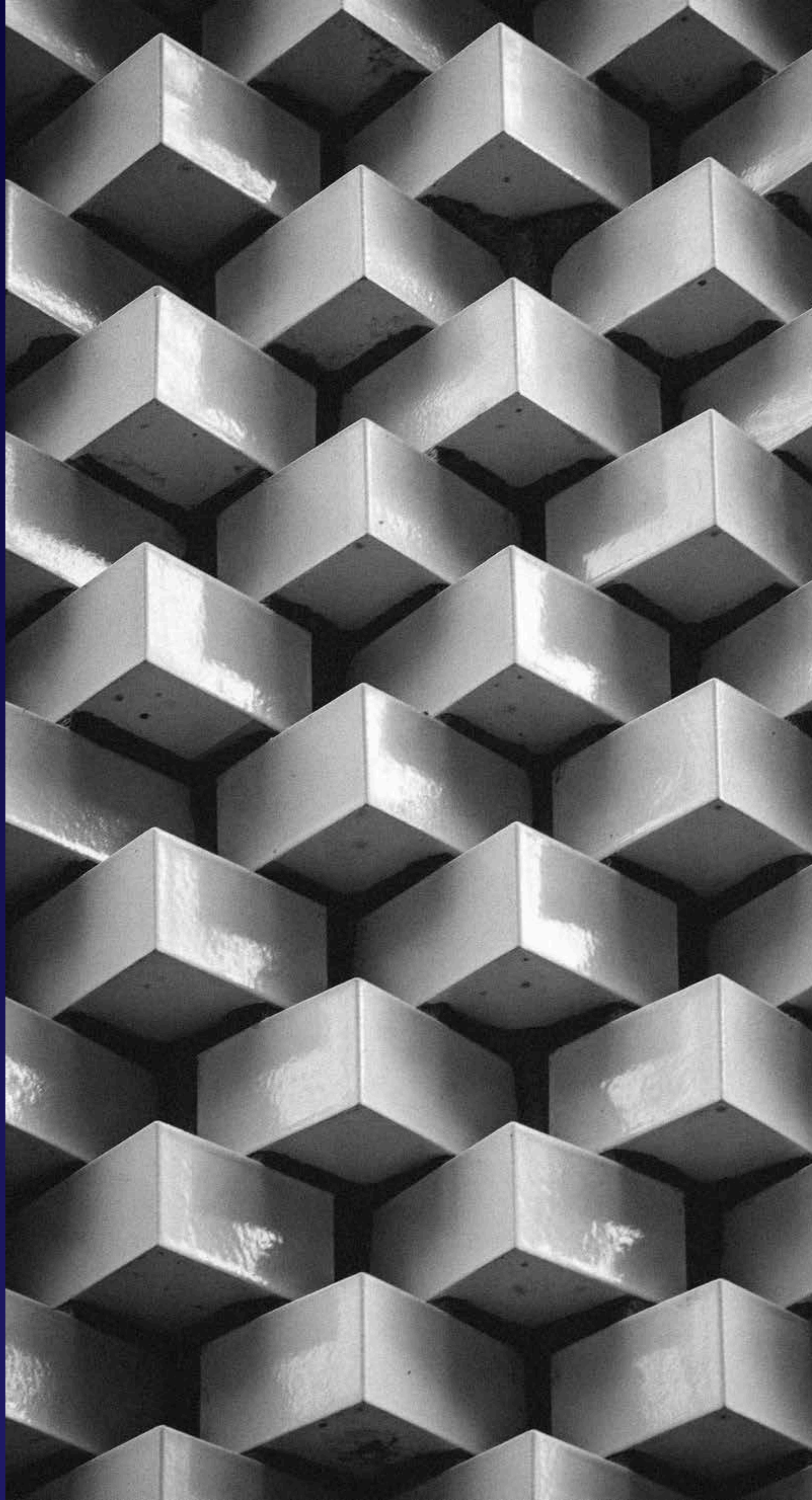
   Management

3. Remember Azure Advisor.

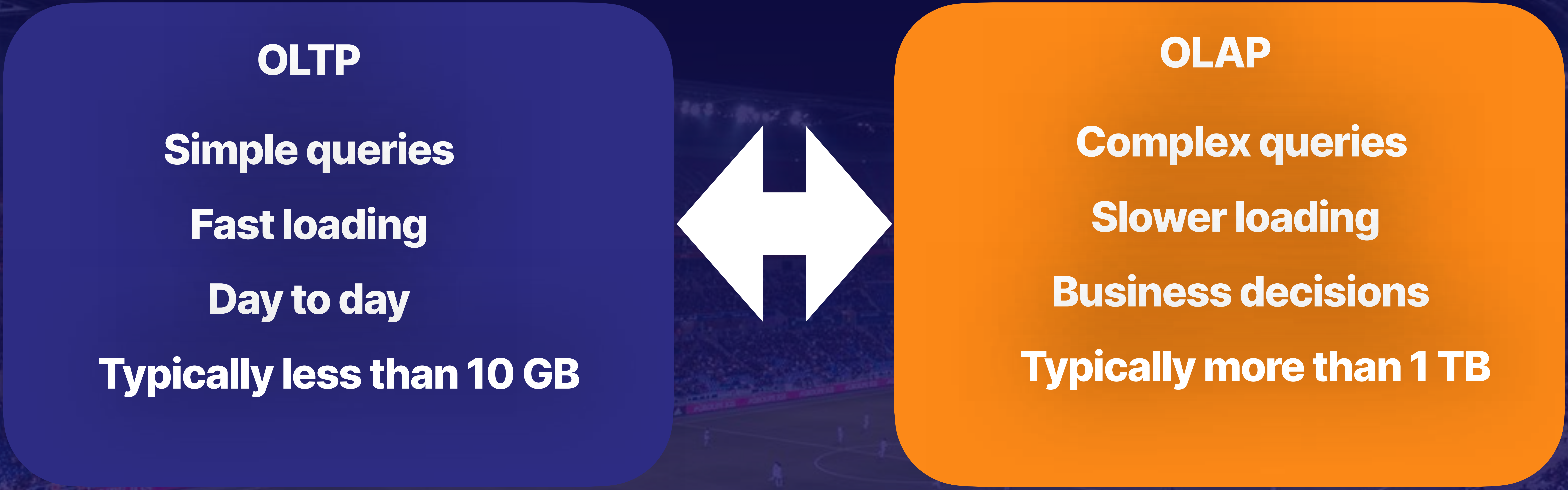
   A must-use

The Basics of Result Set Caching

- SQL pool auto-caches query results in a user database for repeat use.
  - Persisted cache (query performance and less compute)
- You must turn on set caching.
  - User database
  - Session
- What is not cached?
  - User-defined functions
  - Row/column security
  - Rows larger than 64 KB/total data over 10 GB
  - Built-in functions or runtime that isn't deterministic



Online Transaction Processing vs. Online Analytical Processing



OLTP

Simple queries

Fast loading

Day to day

Typically less than 10 GB

OLAP

Complex queries

Slower loading

Business decisions

Typically more than 1 TB

Start with Ambari UI

- Configuration settings
- Cluster health
- Stack and version

Examine the Log Files

- Check **`stderr`** and **`syslog`** files.
- Check Hadoop step logs.

Configuration Settings

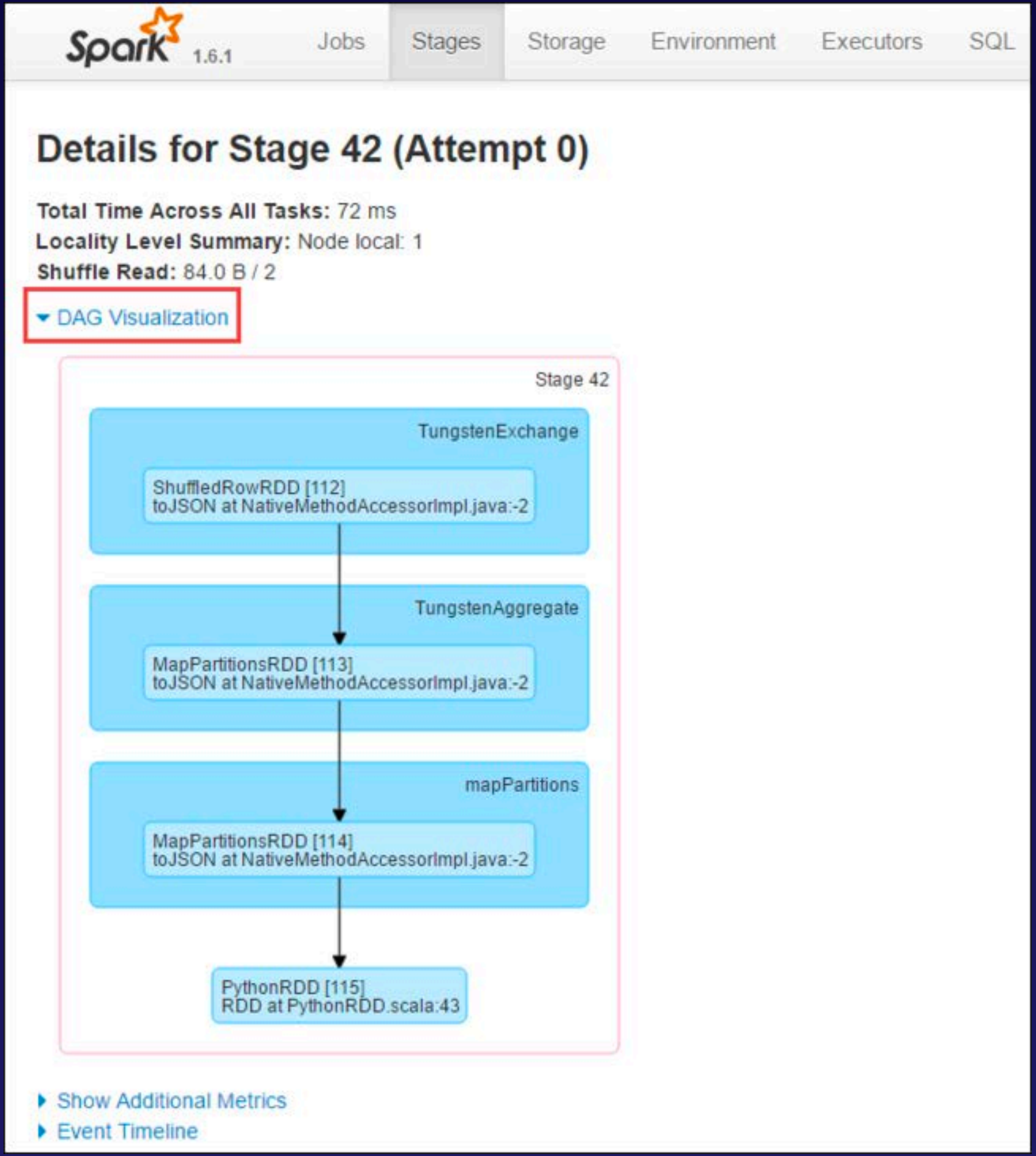
- Have you optimized your configuration settings?
  - Cluster settings
  - Hardware configuration
  - Nodes

Reproduce the Error

- If all else fails, try again on a new cluster.

Tracking Applications in the Spark UI

- Jobs
  - Pull detailed information on submitted jobs
- Executors
  - Broken down by ID
  - Task information
  - Memory and shuffle usage
- Stages
  - Shuffle read/write
  - Duration and I/O
  - See a DAG visualization of each stage



Details for Stage 42 (Attempt 0)

Total Time Across All Tasks: 72 ms
Locality Level Summary: Node local: 1
Shuffle Read: 84.0 B / 2

▾ DAG Visualization

Stage 42
TungstenExchange
ShuffledRowRDD [112]
toJSON at NativeMethodAccessorImpl.java:-2
TungstenAggregate
MapPartitionsRDD [113]
toJSON at NativeMethodAccessorImpl.java:-2
mapPartitions
MapPartitionsRDD [114]
toJSON at NativeMethodAccessorImpl.java:-2
PythonRDD [115]
RDD at PythonRDD.scala:43

▸ Show Additional Metrics
▸ Event Timeline