Developing Batch Processing Solutions with Azure HDInsight

Tim Warner

AUTHOR EVANGELIST, PLURALSIGHT

@TechTrainerTim TechTrainerTim.com

Overview

Understand Hadoop

Implement HDInsight to perform batch processing

- Hive
- Spark

Understand Hadoop

What is Apache Hadoop?

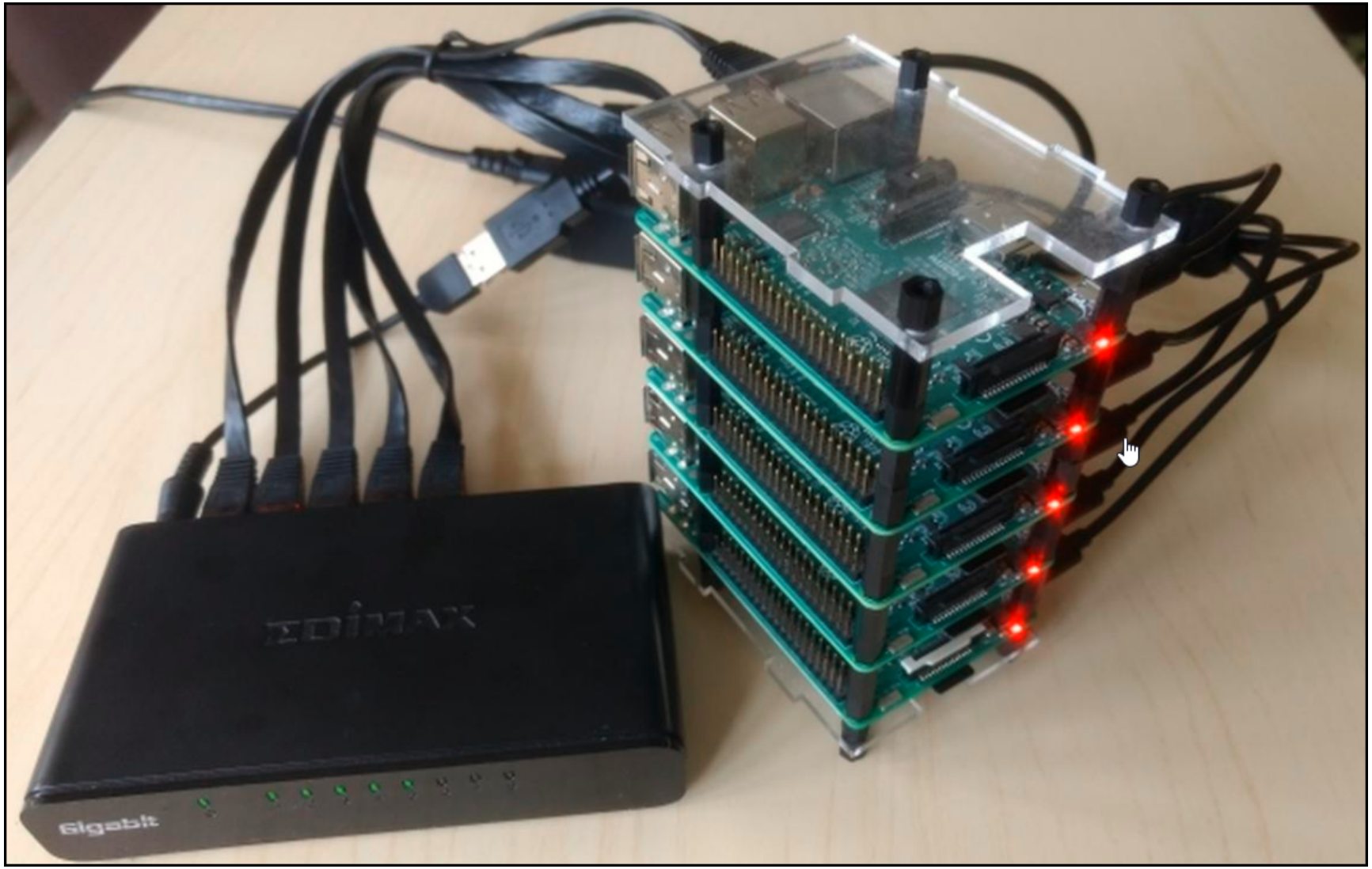
Original open-source framework for distributed big data processing and analysis

Based on Google File System

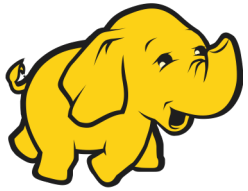
Core components:

- Hadoop File System (HDFS)
- YARN resource manager
- MapReduce processing algorithm
  - Java

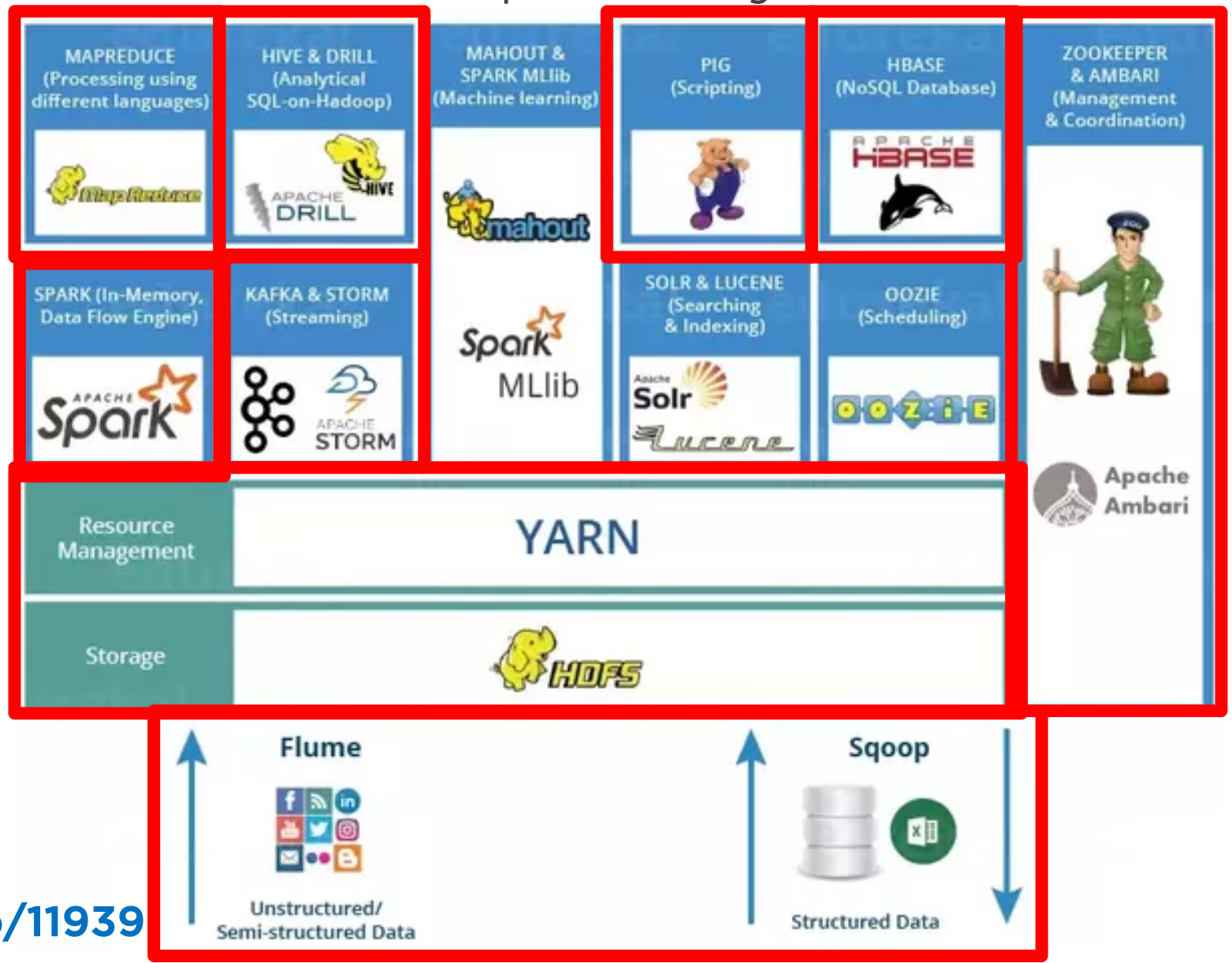
Commodity Hardware



https://timw.info/2a141



Hadoop Ecosystem



https://timw.info/11939

Hadoop vs. Traditional RDBMS

RDBMS

Structured data

ACID

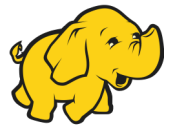
Lower data throughput, but faster granular query performance

Vertically scaled

OLTP

SQL Server/Azure SQL Database are licensed and closed source

Hadoop



Unstructured (although you can project structure) – "schema on read"

CAP

Higher data throughput, but slower granular query performance

Horizontally scaled

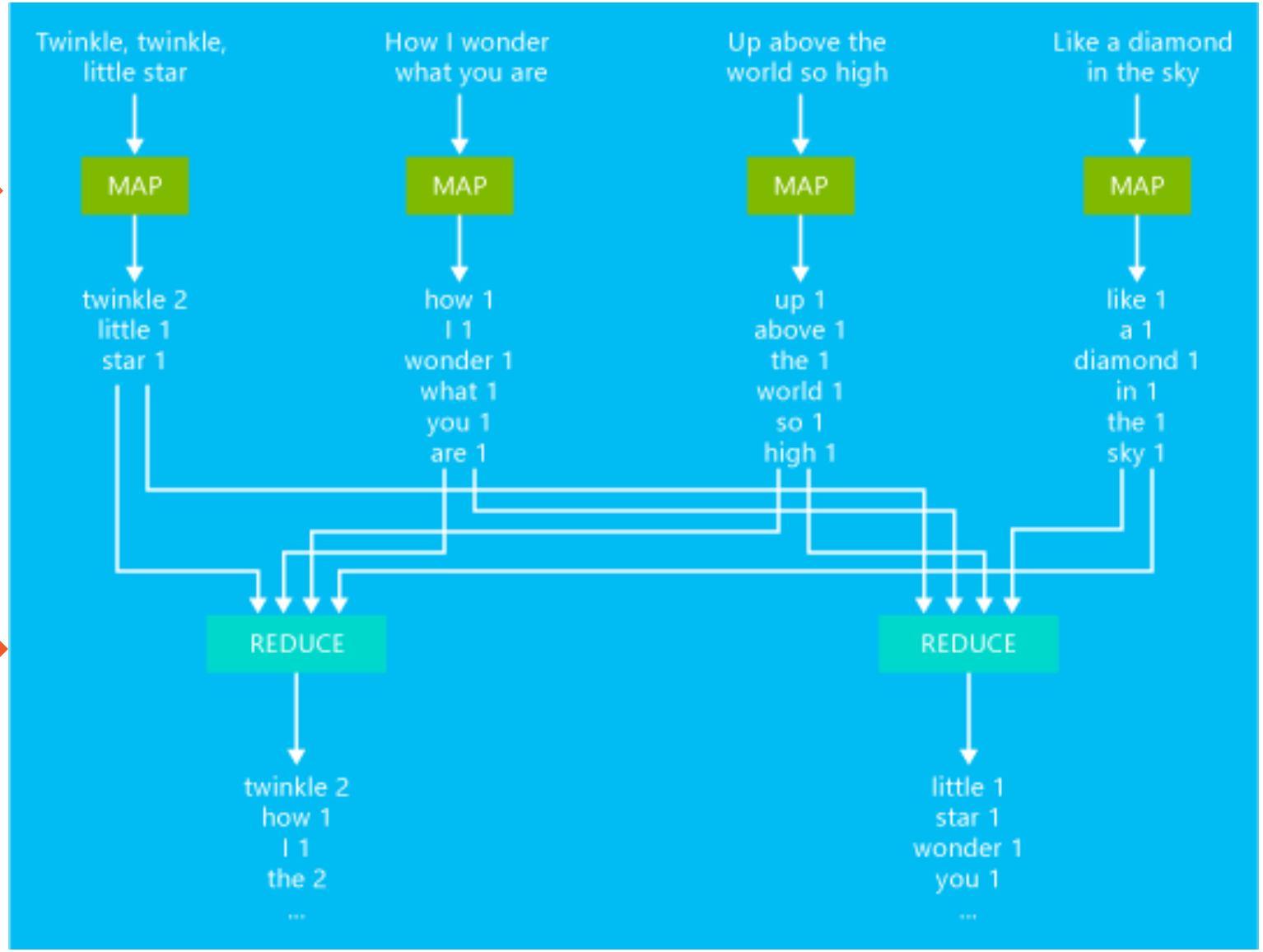
OLAP

(Mostly) free and open source

MapReduce Operation

Data is chunked redundantly across nodes

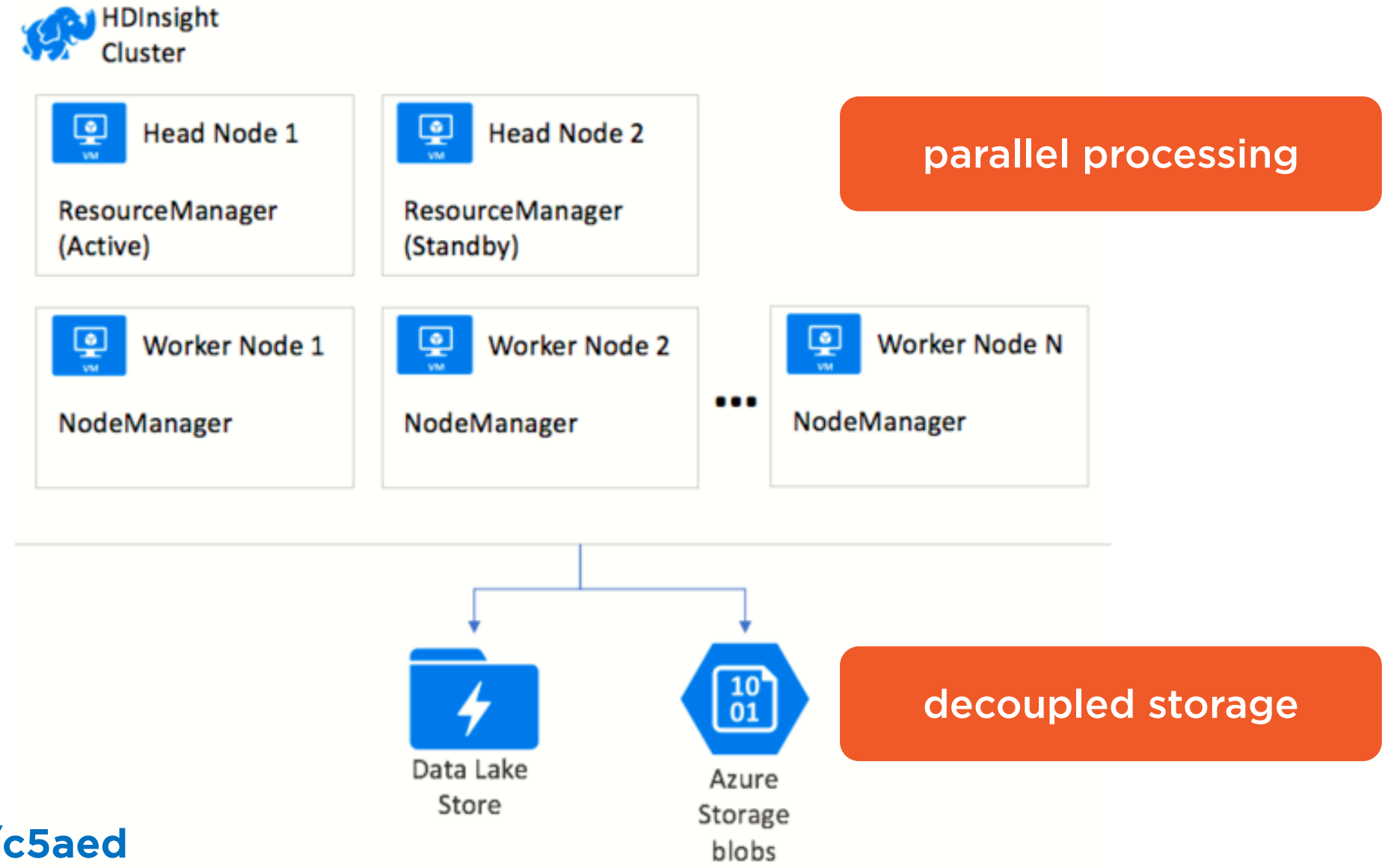
Massive parallelism



https://timw.info/d094

Implement HDInsight

HDInsight High-Level Architecture



https://timw.info/c5aed

HDInsight Cluster Types

Hadoop

- Batch query and analysis of HDFS stored data

HBase

- Processing for large schemaless NoSQL data

Interactive Query

- In-memory caching for fast Hive queries

Kafka

- Distributed streaming data platform

ML Services

- Predictive modeling and machine learning

Spark

- In-memory processing and interactive queries

Storm

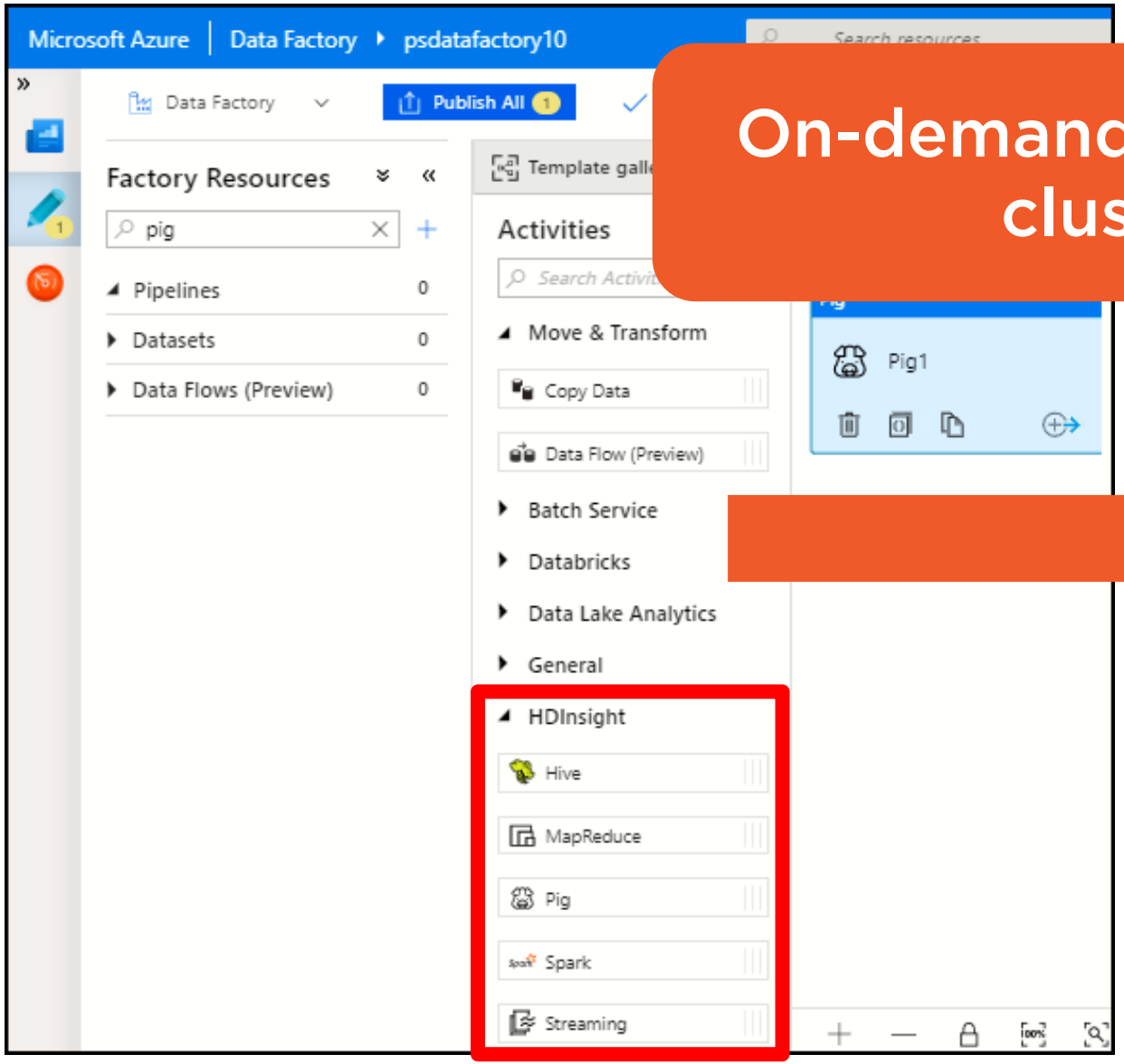
- Real-time event processing

Demo

1

Provision an HDInsight cluster

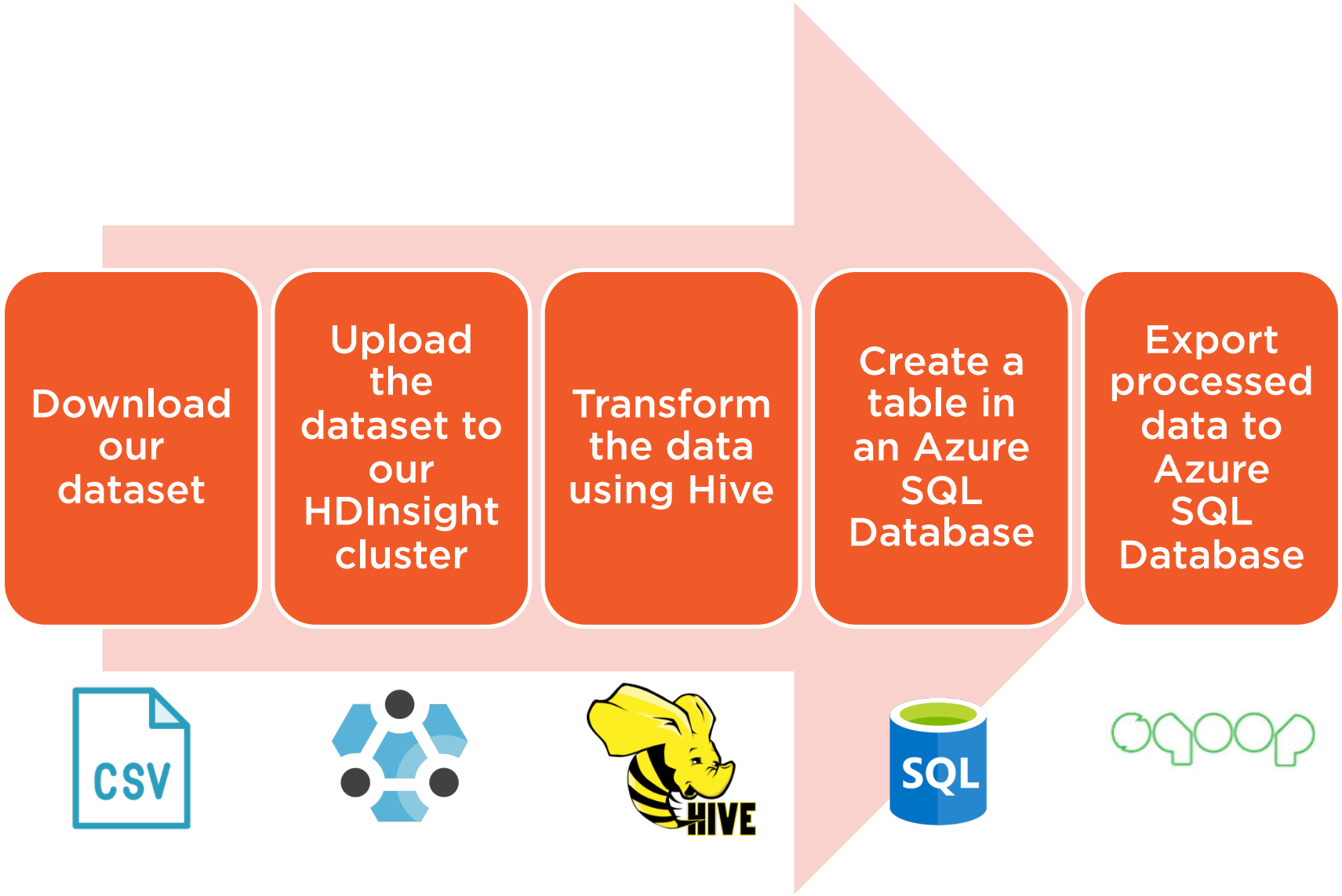
Azure Data Factory Integration



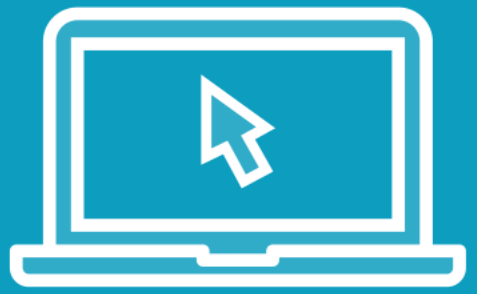
On-demand HDInsight cluster

```
{
                                                    otion",
                                                    yHDInsightLinkedService",
                                                    viceReference"
    },
    "typeProperties": {
        "scriptLinkedService": {
            "referenceName": "MyAzureStorageLinkedService",
            "type": "LinkedServiceReference"
        },
        "scriptPath": "MyAzureStorage\\PigScripts\\MyPigSript.pig",
        "getDebugInfo": "Failure",
        "arguments": [
            "SampleHadoopJobArgument1"
        ],
        "defines": {
            "param1": "param1Value"
        }
    }
}
```

Our HDInsight/Hive Batch Processing Job



Demo



2

Run the tutorial:
https://docs.microsoft.com/en-us/azure/hdinsight/interactive-query/interactive-query-tutorial-analyze-flight-data

About Apache Spark

Processing engine that serves as a MapReduce alternative

Goal: make MapReduce's scale and fault-tolerance faster via in-memory processing

Language support: Scala, Python, Java, R, and SQL

Data science frameworks support: TensorFlow, PyTorch, scikit-learn

Azure Databricks

Ecosystem built around Apache Spark

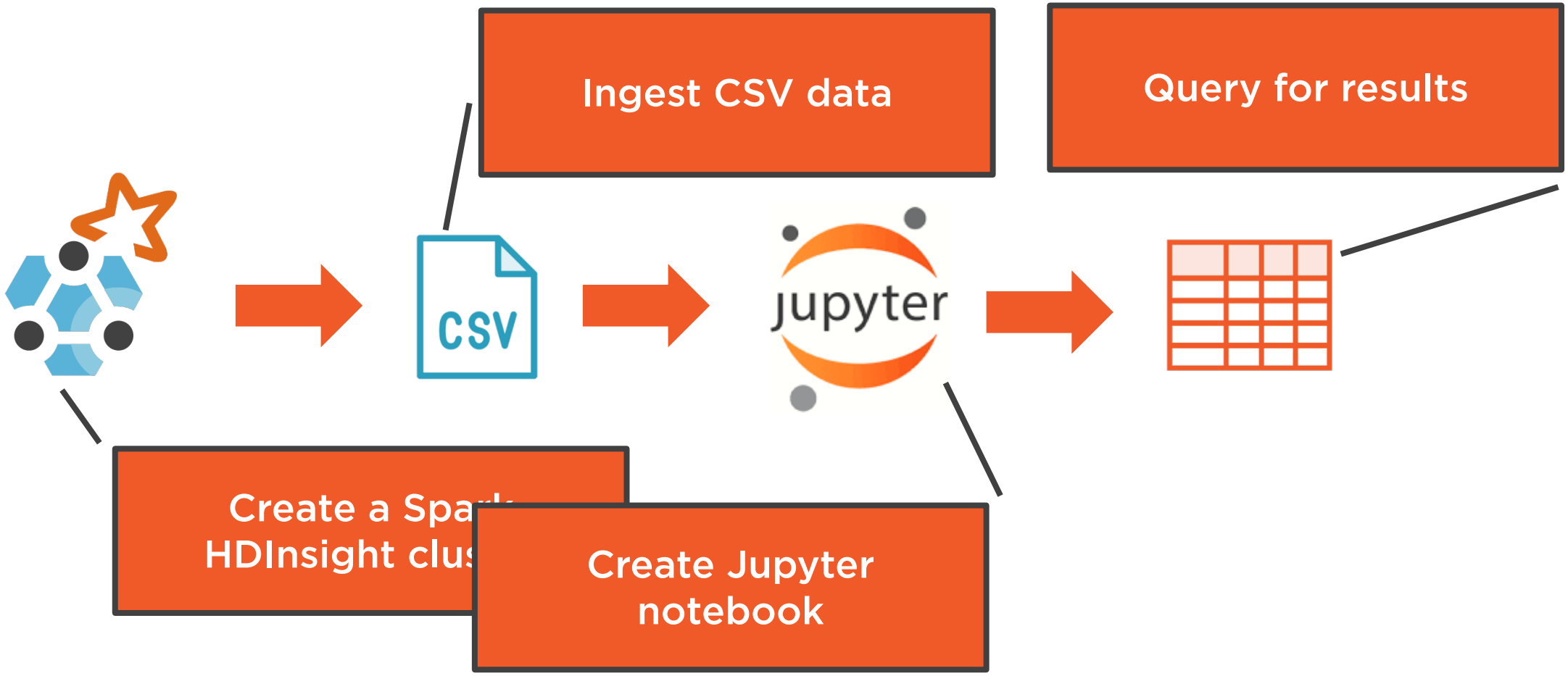
Azure Databricks is Microsoft's hosted environment

Fast, optimized, auto-scaled environment

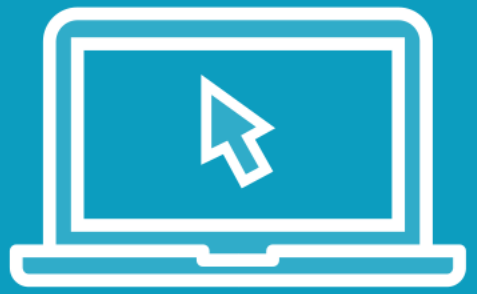
Jupyter notebooks

Integration with Azure ecosystem (Data Factory)

Our HDInsight/Spark Batch Processing Job



Demo



3

Run the tutorial:

https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-load-data-run-query

Next module:

https://docs.microsoft.com/en-us/azure/hdinsight/spark/apache-spark-load-data-run-query

For Further Learning

The Building Blocks of Hadoop
(Janani Ravi)
Taught by a former Google engineer

HDInsight Deep Dive: Storm, HBase, and Hive
(Elton Stoneman)
Employs a real-world scenario

Summary

Microsoft designed HDInsight in conjunction with Hortonworks

- First-class Hadoop experience

Lower compute costs mean Apache Spark is moving to the forefront of big data analysis

Next module: Developing Batch Processing Solutions with Azure Databricks