Enabling Real-Time Analytics for Hadoop Data Lakes with GridGain

Denis Magda,
GridGain, VP of Product Management
Apache Ignite, PMC Chair

Digital Transformation and The Need for Real-time Analytics

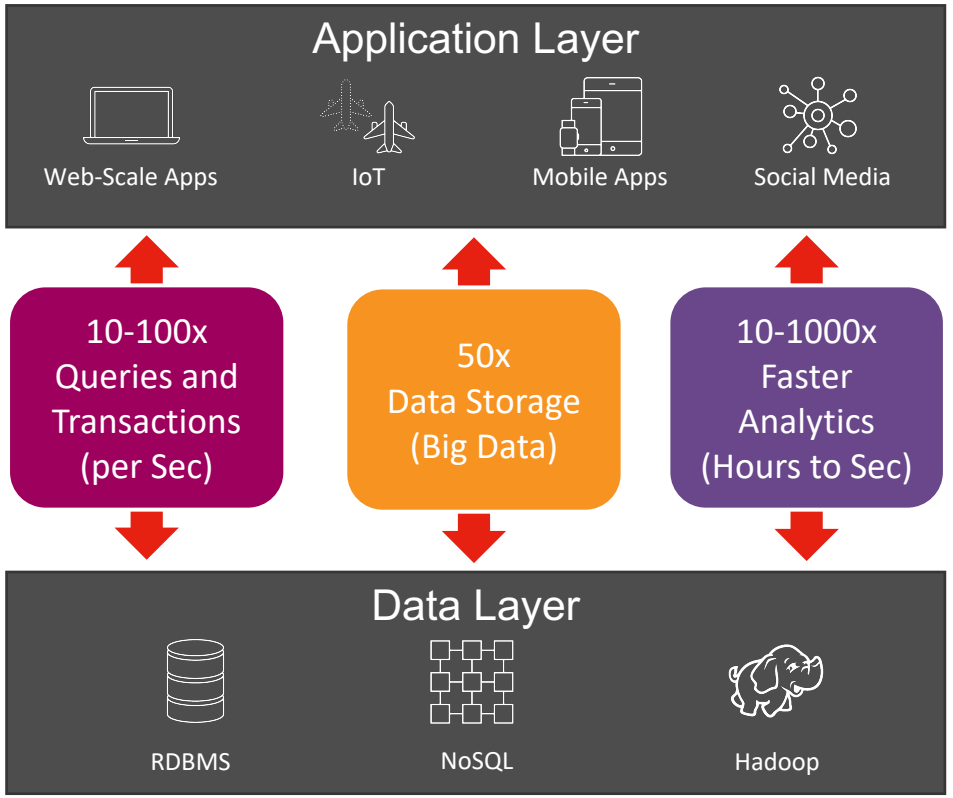
The Common Digital Transformation Goal

"We want to be a tech company with a banking license"

Ralph Hamers, CEO

But Traditional IT Architectures Cannot Support the Modern, Digital Enterprise…

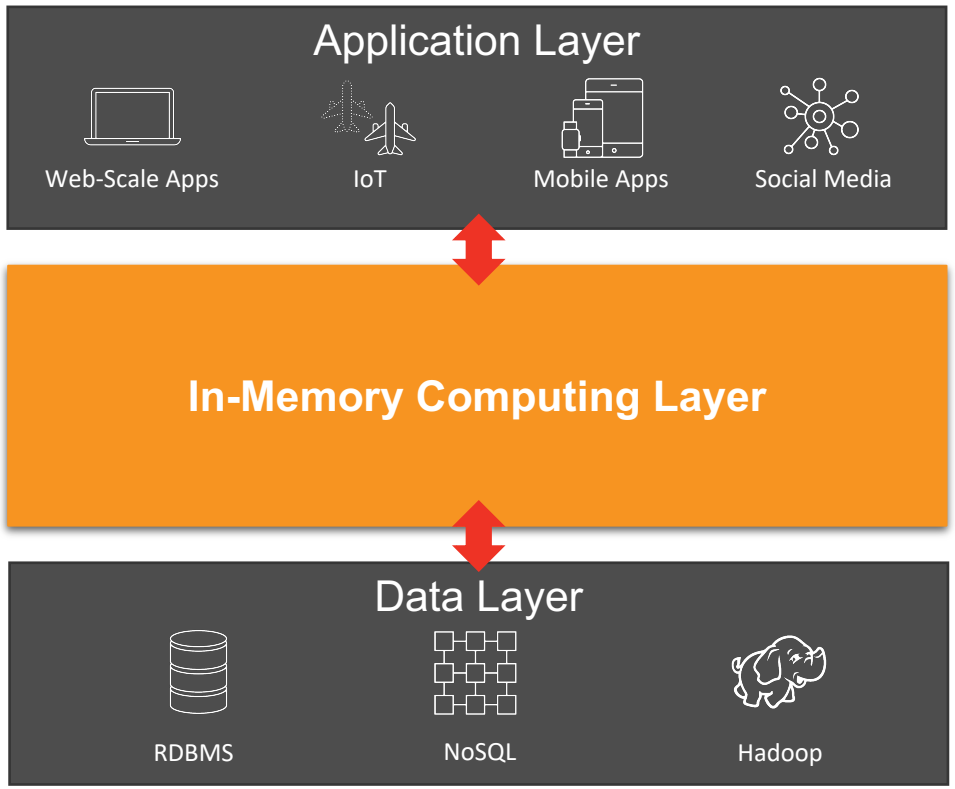
- Digital Transformations Require New Levels of Performance and Scalability
  - 10-100x more queries and transactions per second driven by mobile and digital business
  - 50x as much data today as a decade ago and growing rapidly
  - Overnight analytics must become real-time
- Real-Time Decision Making Requires That Transactions and Analytics are Run On the Same Data Set
  - HTAP, HOAP, Translytical solutions
- Traditional Applications Built on Disk-Based Databases Cannot Provide the Needed Speed and Scalability



GridGain

In-Memory Computing Solves the Speed and Scalability Challenges of Digital Business

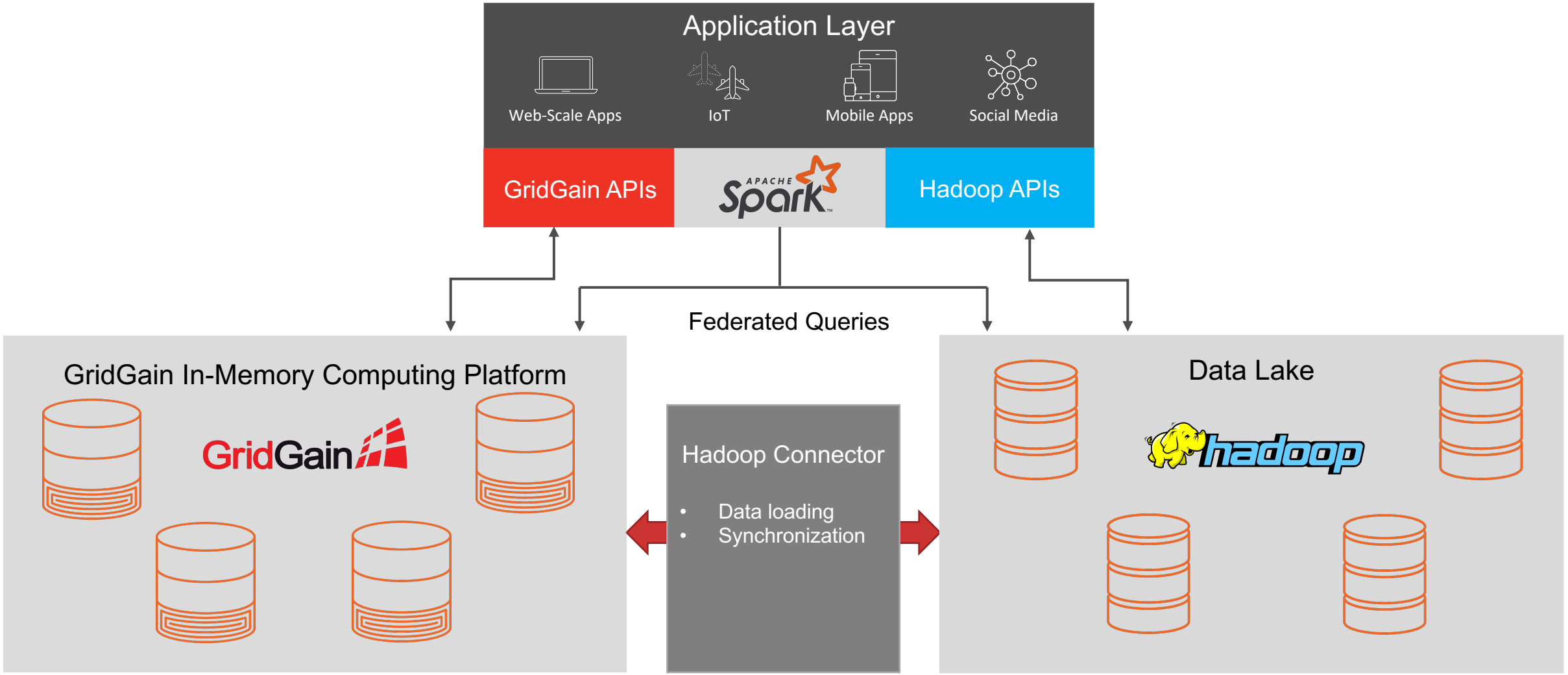
- The Only Affordable Path to Dramatic Speed and Scalability Improvements for Existing Applications
- Application Performance Increases 10x to 1,000x
- Applications Become Scalable to Petabytes of Data
- Operational Data Can be Analyzed in Place in Real-Time Using HTAP/HOAP/Translytical Data Stores



GridGain

Data Lake Acceleration Architecture

Data Lake Acceleration Solution With GridGain



GridGain

Roles of Hadoop and GridGain

Continue Using Hadoop…

- As a storage for old historical data (weeks, months, years)
- For batch processing (dozens of seconds, hours)
- Standard analytics workloads

Switch to GridGain…

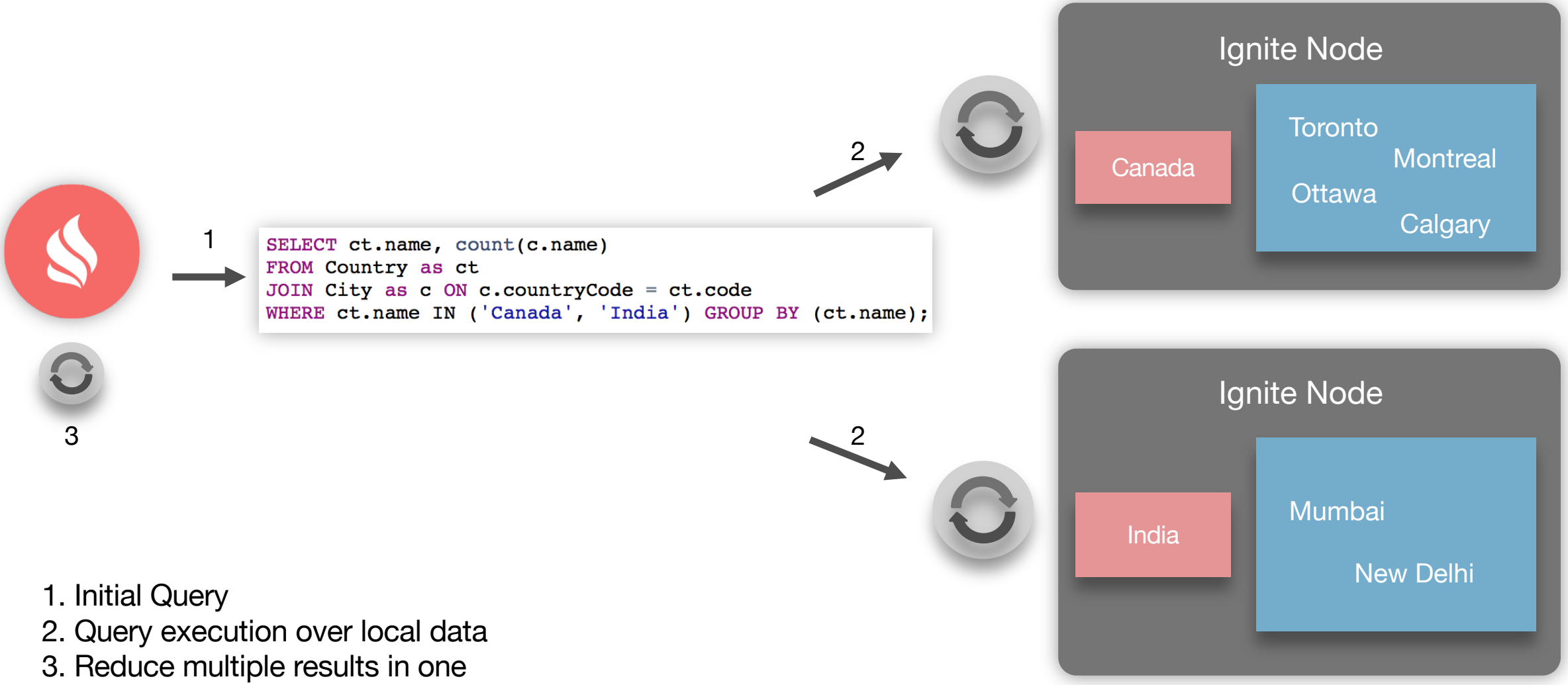
- As a storage for operational and “warm” historical data
- For real-time processing (seconds, millisecond)
- Operational workloads + real-time analytics

Use Spark…

- To span across operational and historical data sets

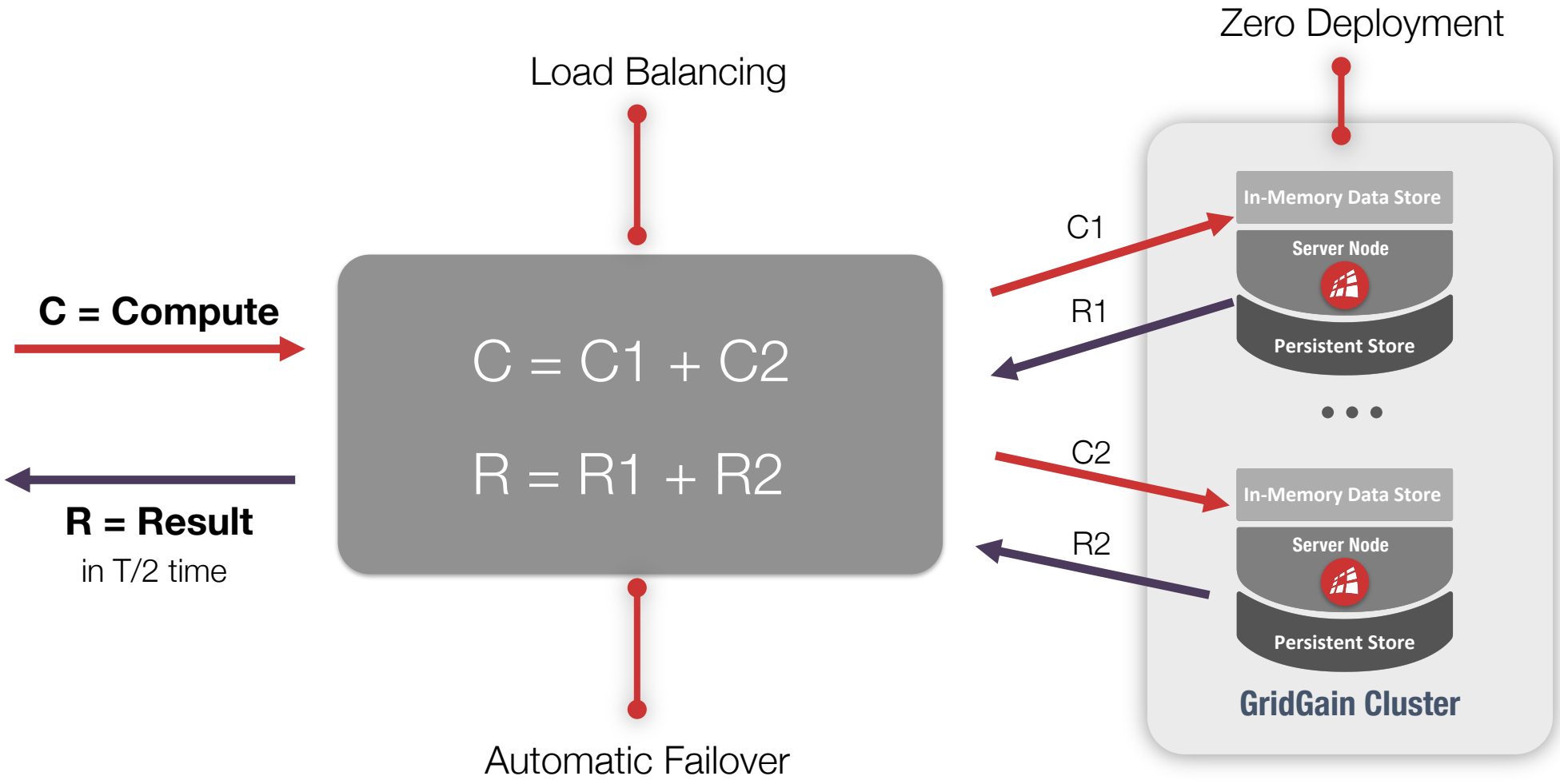
Essential GridGain APIs

Ignite SQL Queries



1. Initial Query
2. Query execution over local data
3. Reduce multiple results in one

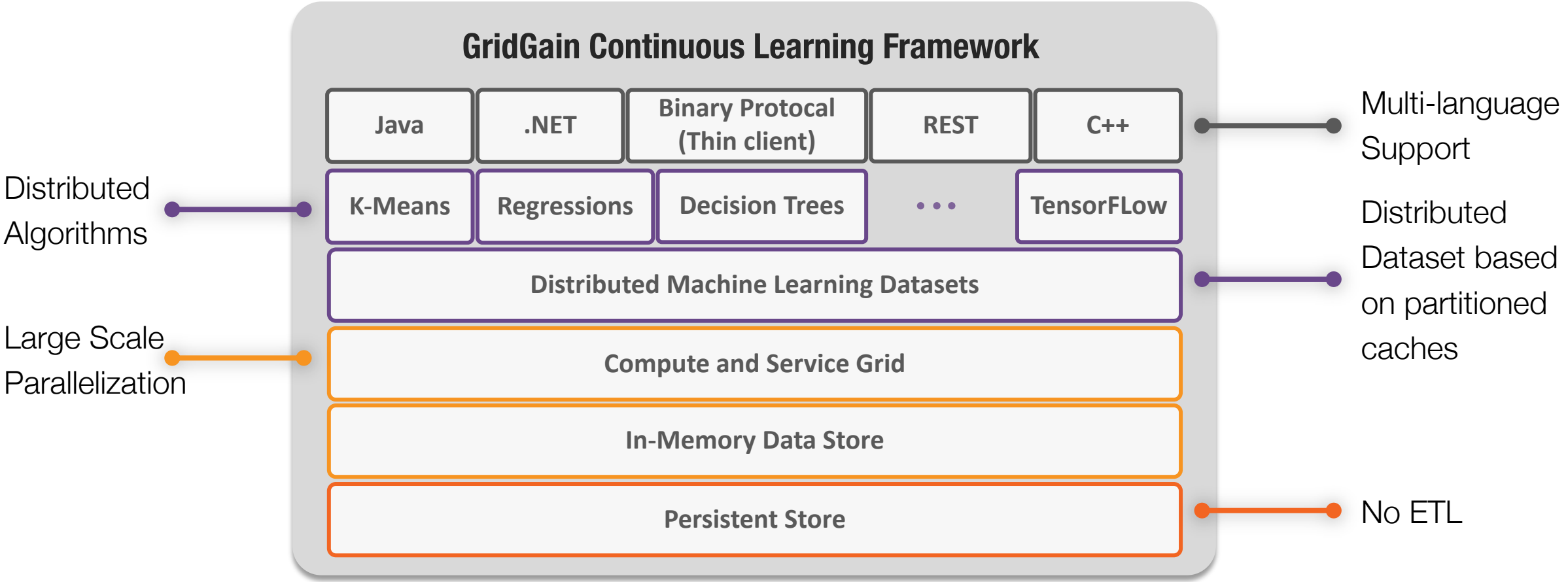
Ignite Compute Grid



GridGain

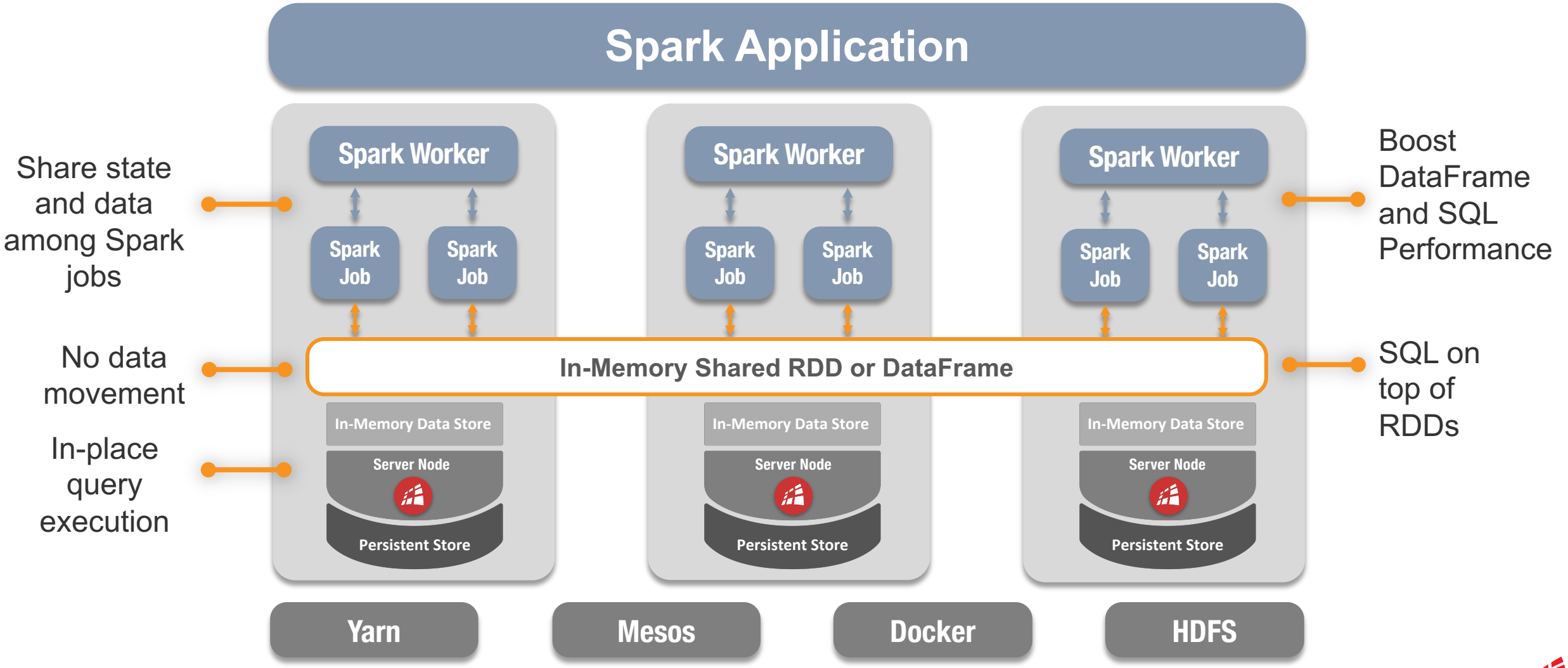
Continuous Learning Framework

In-Memory Machine and Deep Learning



GridGain

GridGain Spark Integration

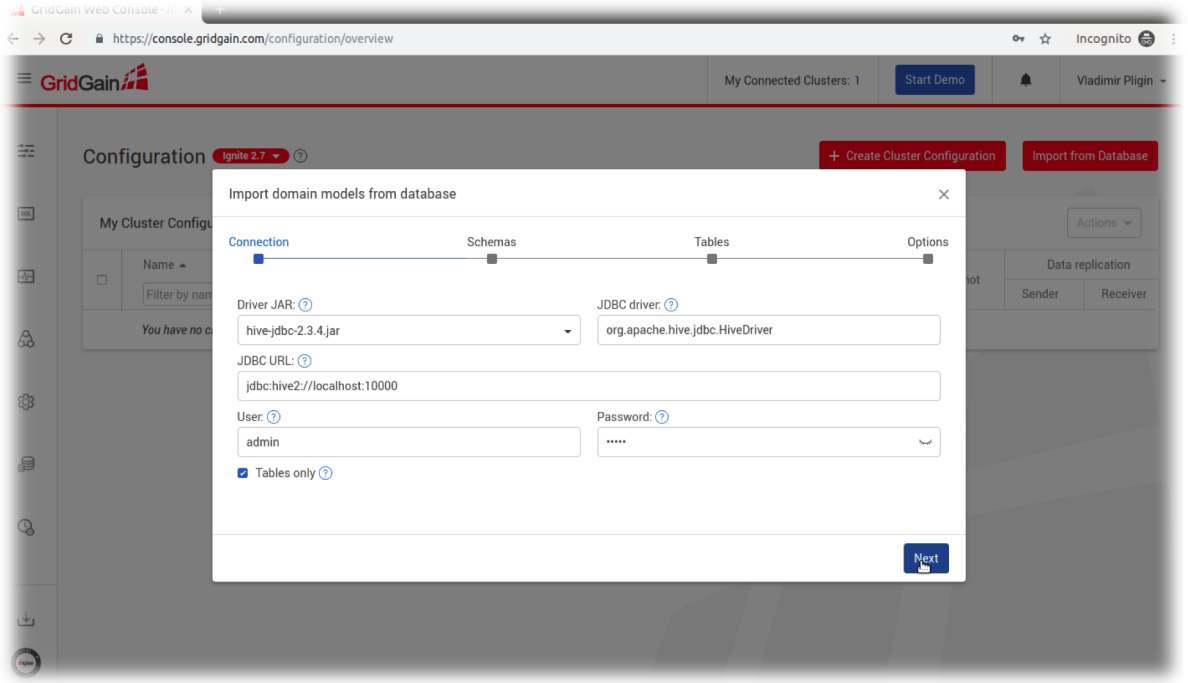


GridGain

GridGain Hadoop Connector

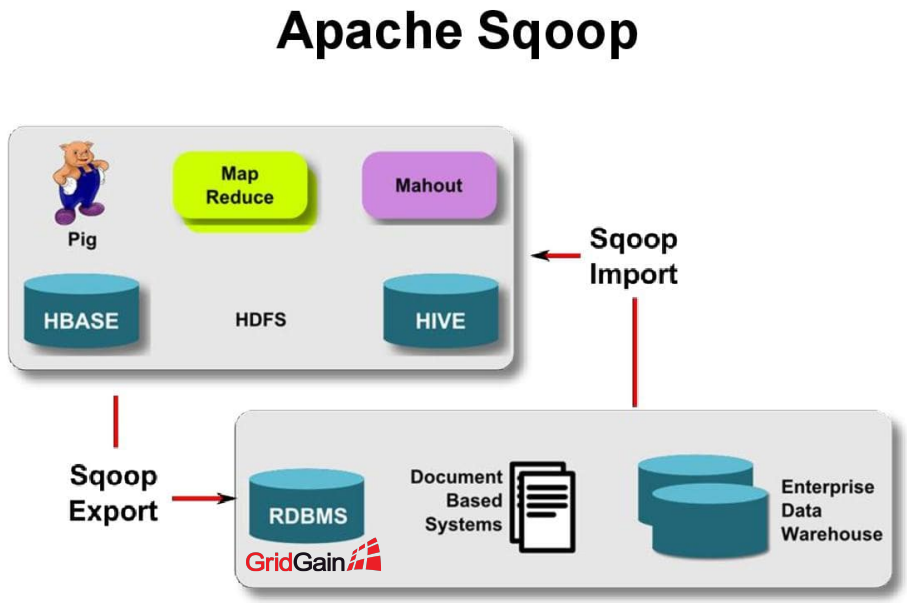
Data Loading

- SparkLoader Tool
  - For deployments with Spark
- Hive Store
  - Schema Import and Loading
  - Writes GridGain changes to Hadoop



Clusters Synchronization with Apache Sqoop

- Keep GridGain and Hadoop in Sync
  - Asynchronous import/export
- Uses JDBC Driver for GridGain
- Captures Incremental Changes
  - Sqoop monitors Timestamp fields for



Federated-Queries

```
//Create Hive DataFrame
Dataset<Row> hiveDS = session.table("default.cities") .select("city_id", "city_name");


//Create GridGain DataFrame
Dataset<Row> gridgainDS = session.read()
.format(IgniteDataFrameSettings.FORMAT_IGNITE())
.option(IgniteDataFrameSettings.OPTION_TABLE(), "Person")
.load().select("id", "city_id", "name", "age", "company");


//INNER JOIN
hiveDS.join(gridgainDS, hiveDS.col("city_id").equalTo(gridgainDS.col("city_id"))).show();
```

Download and Implement Today

- GridGain Data Lake Accelerator Solution
  - https://docs.gridgain.com/docs/bdb-getting-started
- GridGain Hadoop Connector Downloads
  - https://www.gridgain.com/resources/download#bigdatapack

Q&A