Microsoft | Innovation Center Torino

CloudTech

Marrakesh 2016

Microsoft | Innovation Center Torino

CloudTech

Marrakesh 2016

Introducing Microsoft Azure Big Data Platform

https://git.io/vrwyo

Francesco Scullino

Agenda:

Big Data overview
Hadoop
Azure HDInsight
Demo
Hands-on Lab

CloudTech

Marrakesh 2016

Big Data overview

What is Big Data?

- Many definitions:
  - Big data is high **volume**, high **velocity**, and/or high **variety** information assets that require new forms of processing to enable enhanced decision making, insight discovery and process optimization Gartner 2012
  - ….
  - Big Data happens when the data you have to process is bigger than what you can process in the given time with current technologies Silicon Angle

- Big Data is often described using 3/4/5 V

Volume

- Refer to the **large amount** of data to be stored and analyzed

  - Traditional storage have upper memory limits

Velocity

- Refer to the **speed** at which new data are generated shared

- It needs technologies to analyze data in near real time or real time

Variety

- Refer to the **different types** of data we can process

Structured

CUSTOMER

NAME	DATATYPE	NULLABLE?
CUSTOMER_ID	VARCHAR	NO
FIRST_NAME	VARCHAR	NO
LAST_NAME	VARCHAR	NO
BIRTH_DAY	TIMESTAMP	NO
ADDRESS	VARCHAR	NO
ADDRESS2	VARCHAR	YES
STATE	VARCHAR	NO
ZIP_CODE	INTEGER	NO

CUST_ORDER

NAME	DATATYPE	NULLABLE?
ORDER_ID	VARCHAR	NO
CUSTOMER_ID	VARCHAR	NO
STATUS	VARCHAR	NO
ORDER_AMOUNT	DECIMAL	NO

PRODUCT

NAME	DATATYPE	NULLABLE?
PRODUCT_ID	VARCHAR	NO
CATEGORY	VARCHAR	NO
LIST_PRICE	DECIMAL	NO

Unstructured

Semi-structured

```
{"menu": {
  "id": "file",
  "value": "File",
  "popup": {
    "menuitem": [
      {"value": "New", "onclick": "CreateNewDoc()"},
      {"value": "Open", "onclick": "OpenDoc()"},
      {"value": "Close", "onclick": "CloseDoc()"}
    ]
  }
}}
0123456789
```

… and

[Veracity]: refer to the **quality** of the data

[Value]: refer to the **business** decisions that can be taken at the end of big data processing

Data sources

- Conversations
- Social Media
- Application logs
- Photos
- Videos
- Text
- …

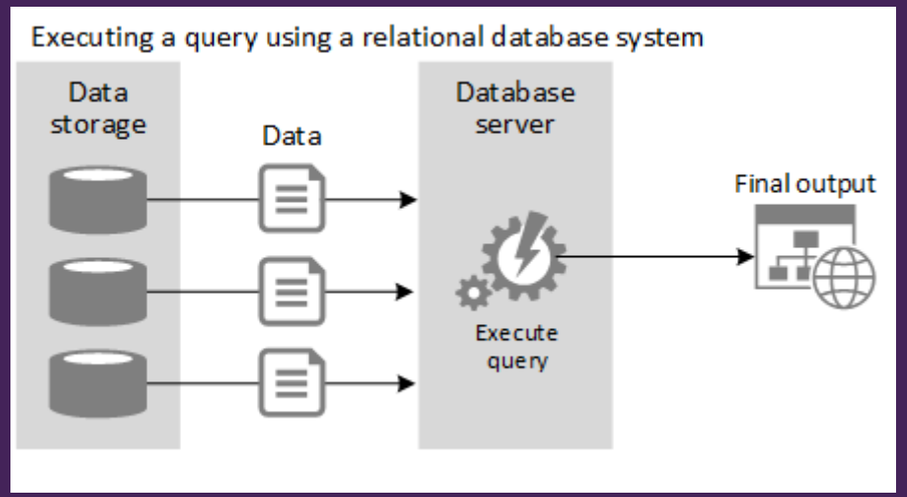
Value

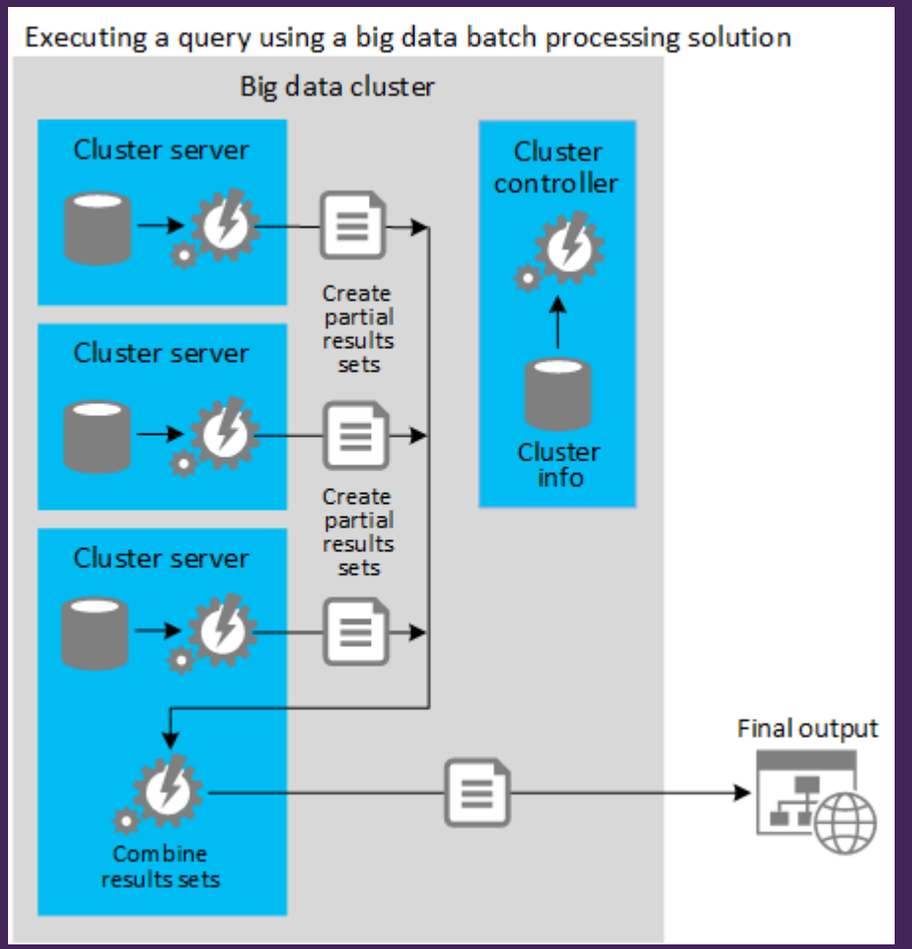
- Understand and target customers analyzing web search trends
- Disaster management analyzing social media
- Improve security analyzing conversation
- Internet of things: store and analyze data coming from a sensor network
- …

How do big data solutions work?

- Big data batch breaks up source data files into multiple blocks and replicates the blocks on a distributed cluster of nodes (servers). Data processing runs in parallel on each node, and the parallel processes are then combined into an aggregated result set.

RDBMS Query vs Big Data Processing





CloudTech

Marrakesh 2016

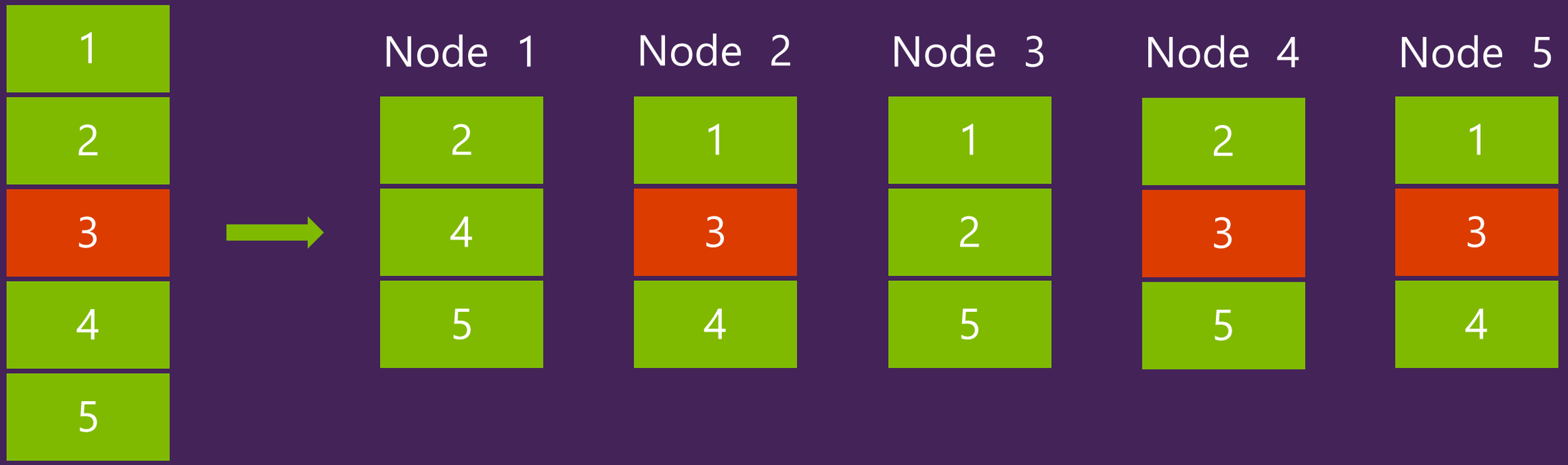
Hadoop

Hadoop

At the core of many big data implementations is an open source framework named Apache Hadoop

- Core modules
  - **HDFS**
  - **MapReduce**: framework to compute distributed processing using Java language
  - Common utilities
  - **YARN**: job scheduler and cluster manager

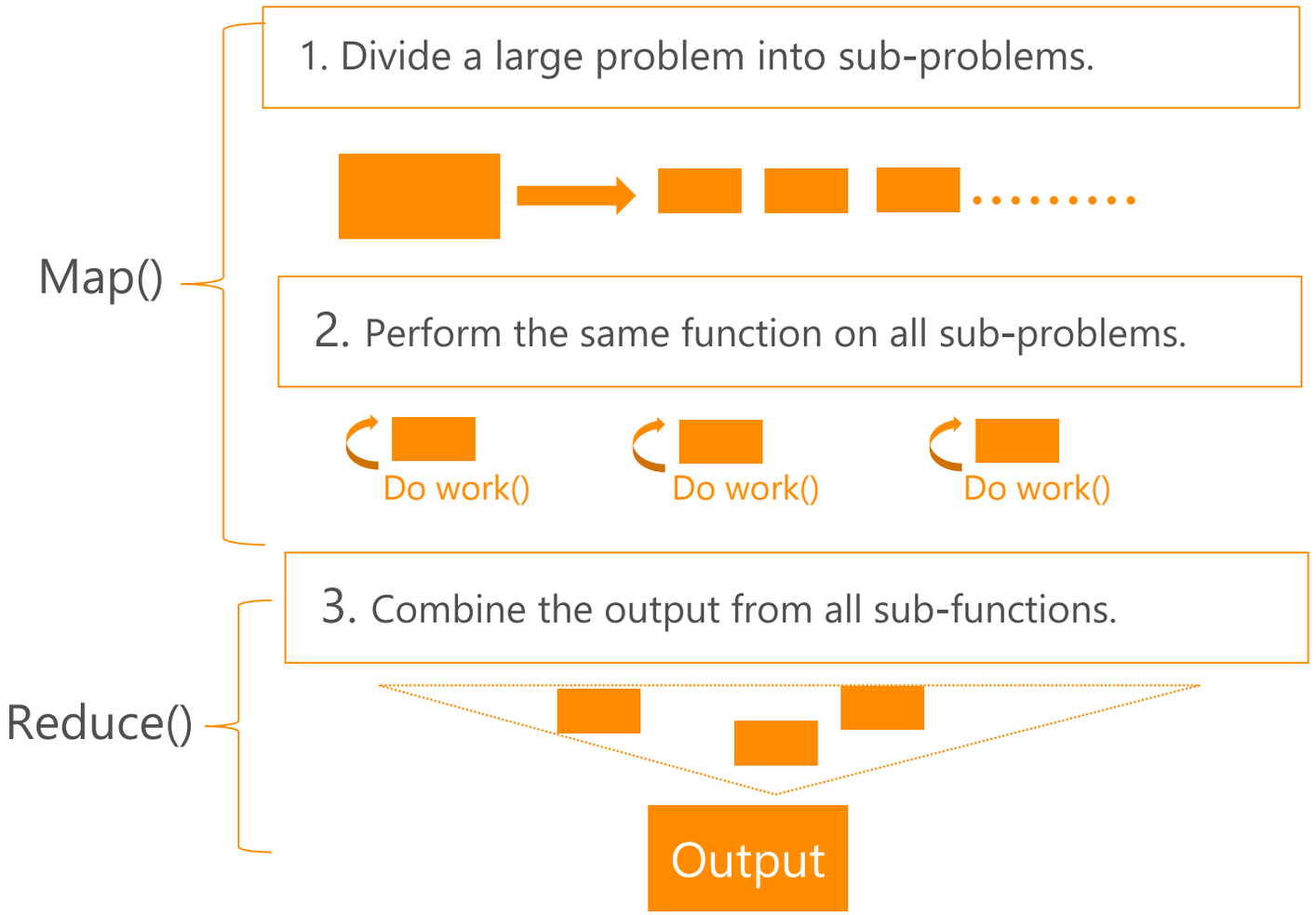
HDFS data distrubution



Input file

Hadoop MapReduce

- Programming framework (library and runtime) for analyzing datasets stored in HDFS
- Composed of user-supplied Map and Reduce functions:
  - Map() - subdivide and conquer
  - Reduce() - combine and reduce cardinality

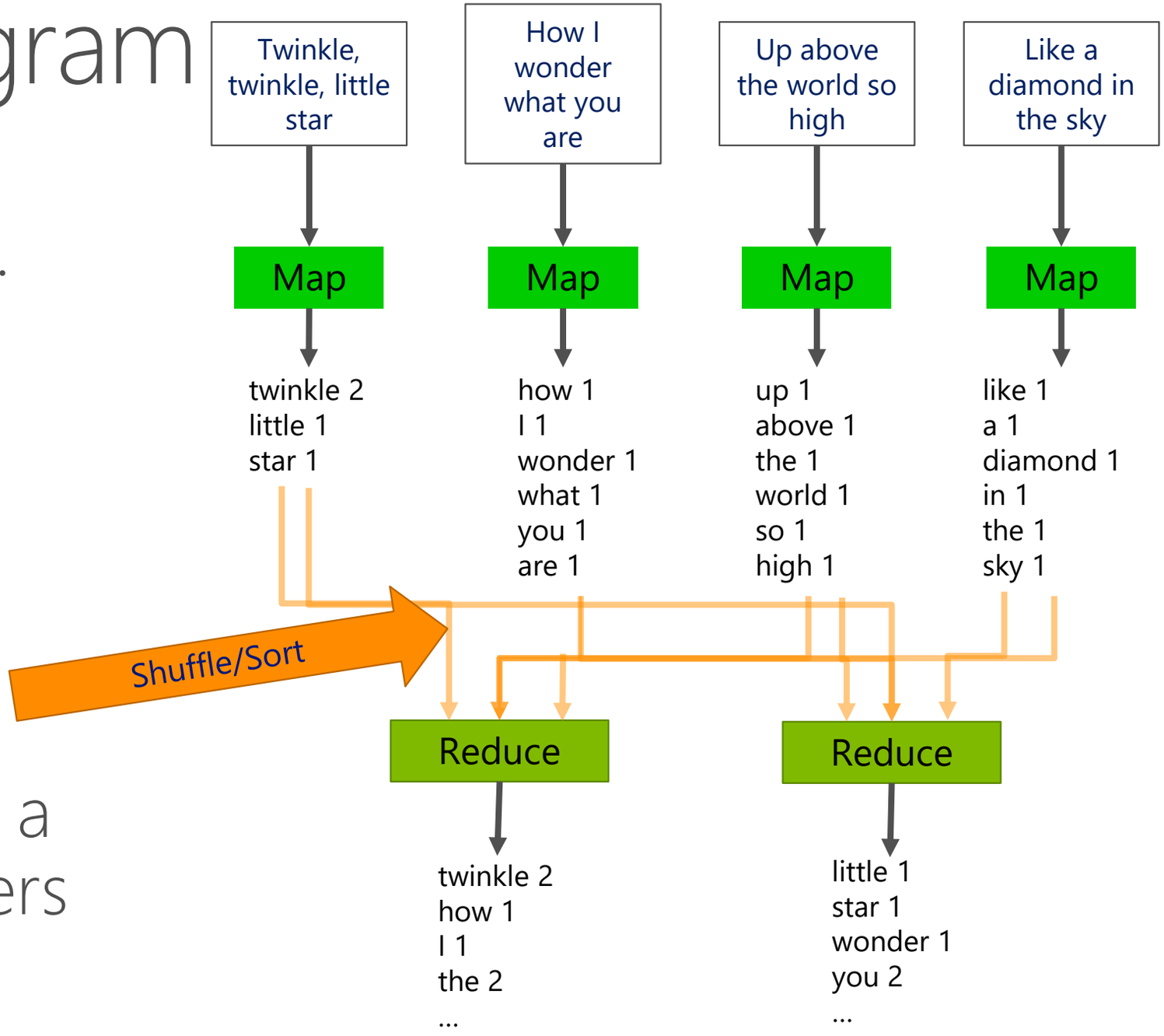


MapReduce program

Programs = Sequence of "map" and "reduce" tasks.

Process large volumes of data in parallel

Divides the work into independent tasks across a large number of computers



Hadoop

- Other tools
  - **Ambari**: A web-based tool for provisioning, managing, and monitoring Apache Hadoop clusters
  - **HBase**: NoSQL column family database for large tables
  - **Hive**: A data warehouse infrastructure that provides data summarization and ad hoc querying
  - **Mahout**: A scalable machine learning and data mining library
  - **Pig**: A high-level data-flow language and execution framework for parallel computation
  - **Storm**: distributed real time computation system for processing large stream of data
  - **Spark**: process data in memory
  - …

CloudTech

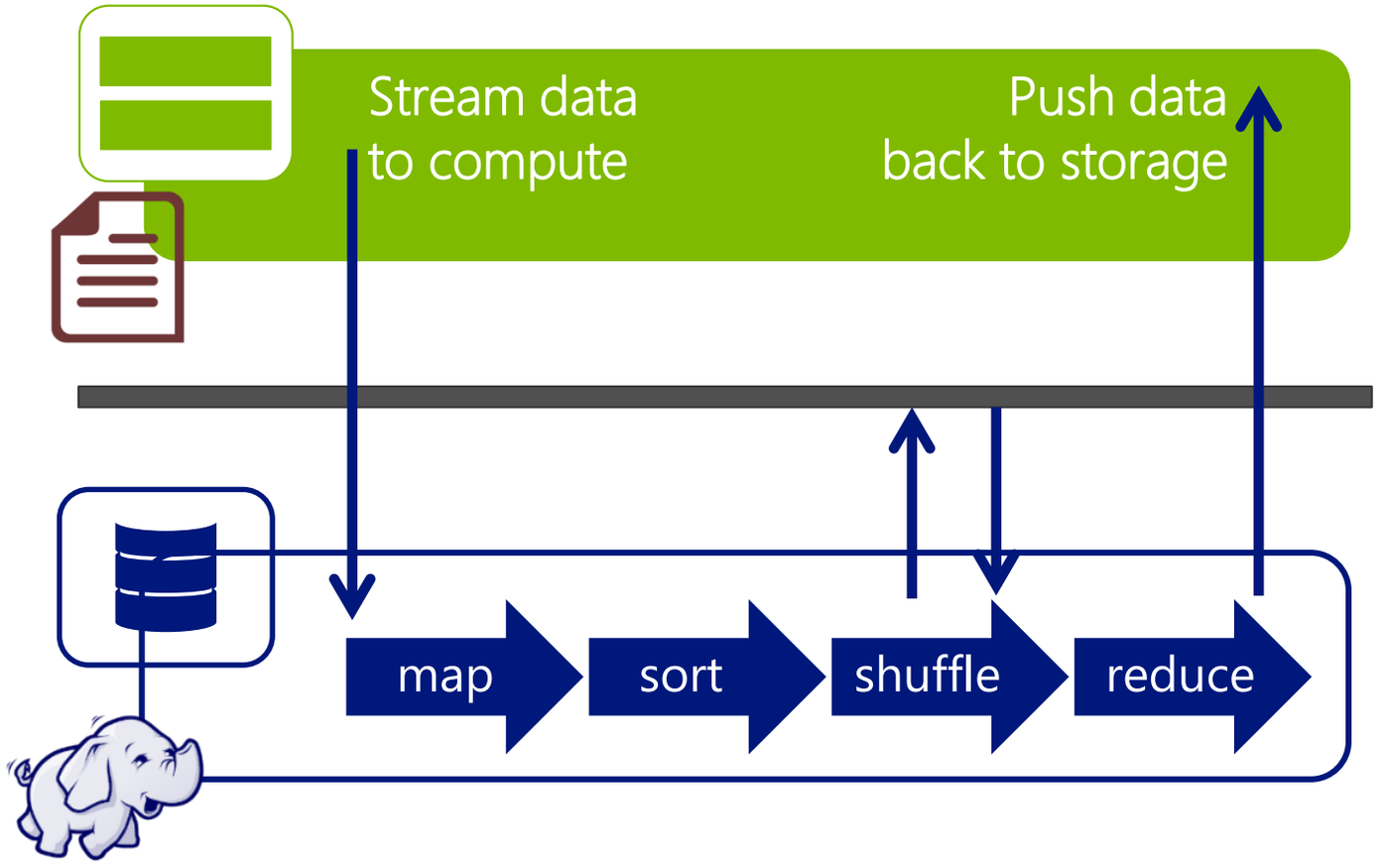
Marrakesh 2016

Azure HDInsight

Hadoop on Azure = HDInsight

- Windows (Hortonworks Data Platform) and Linux clusters
- Supports customization through RDP/SSH and/or Script Action
  - Install additional components like Spark, R, Kafka, Solr, Girafe or customize settings
- Languages: Java, C#, Python, F#
- Separate Compute from Storage
  - Azure Blob Storage in lieu of traditional HDFS

HDInsight Storage Infrastructure



Azure Blob Storage

Azure Flat Network Storage

HDInsight Compute Nodes
(Large VMs)

CloudTech

Marrakesh 2016

Demo: Wordcount using Java MapReduce

https://git.io/vr29x

Refining Data in Hadoop

Microsoft

Data Preparation with Hive & Pig

Create structure over files

Process and refine data with SQL syntax

Generates/runs MapReduce

“Data Warehouse” focused

Process & shape data

Scripting language for ETL/ELT

Generates/runs MapReduce

Apache Hive

Project structure onto data

Schema on Read

Query data using a SQL-like language called HiveQL

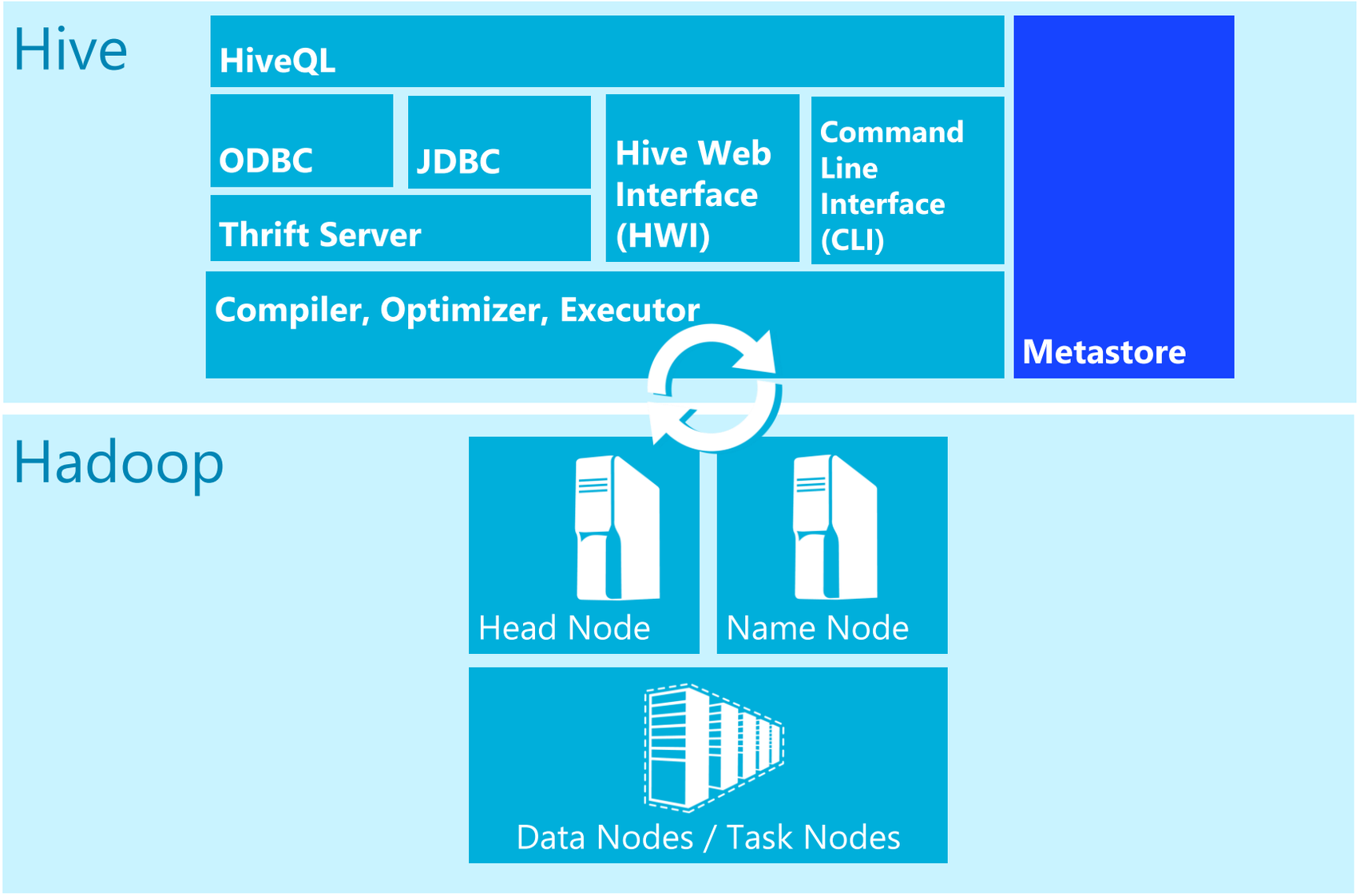
SQL–Like Interface

No Java Needed!

Ad-hoc queries via HiveQL
(translate into MapReduce)

Connect to Microsoft BI and Excel via Hive ODBC

Hive Architecture



Hive

```
SELECT
    get_json_object(json_text,'$.sid') as sid,
    get_json_object(json_text,'$.inc') as inc,
    get_json_object(json_text,'$.status') as status,
    event
FROM bi.event_log
WHERE project='mobile-ios'
    AND dt=20120530
    AND get_json_object(json_text,'$.v')<>'1.5'
    AND (event = 'api_error' OR event = 'api_timeout')
ORDER BY sid;
```

Data Preparation with Hive

External and Internal Tables

```
CREATE EXTERNAL TABLE flights(…column definitions…)
      fields terminated by ','
      lines terminated by '\n'
      stored as textfile
      location 'wasb://cluster.blob.core.windows.net/flights_raw';
```

- Use EXTERNAL when
  - Data used outside Hive
  - You need data to be updatable in real time
  - Data needed when you drop the cluster or the table
  - Hive should not own data and control settings, dirs, etc.
- Use INTERNAL when
  - You want Hive to manage the data and storage
  - Short term usage
  - Creating table based on existing table (AS SELECT)

Apache Pig

Apache Pig is a simple-to-understand data flow language used in the analysis of large data sets.

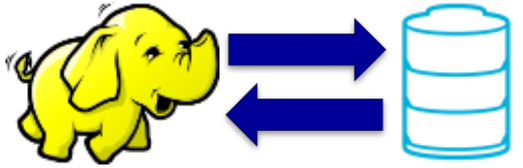
Load: Read data to be manipulated from the file system
Transform: Manipulate the data
Dump or store: Output data to the screen or store for processing

Scripting – No Java Needed!

Pig scripts are automatically converted into MapReduce jobs

Pig's language layer consists of a textual language called Pig Latin and a command shell Grunt

Sqoop



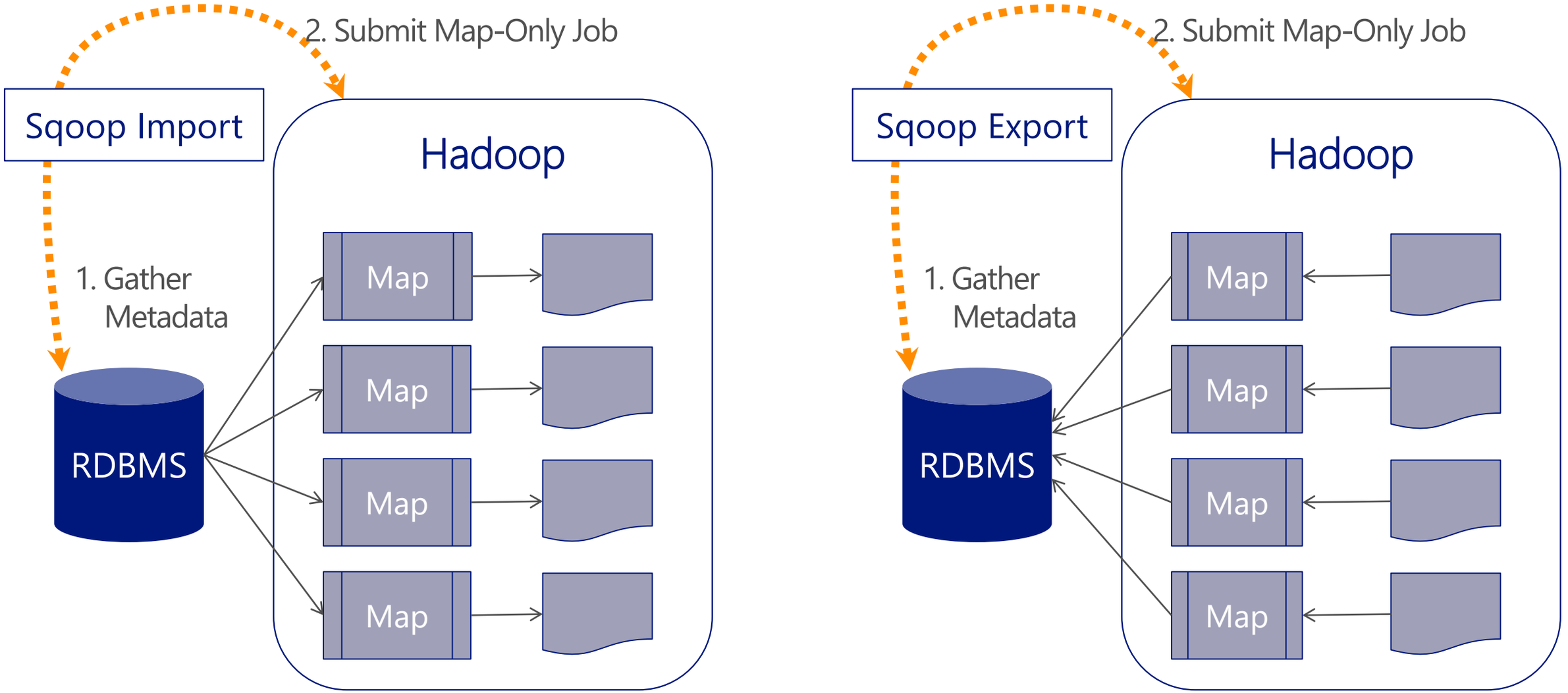
Data connector system for Hadoop and RDBMS

Importing RDBMS data to files (delimited or sequence) in HDFS, or tables in Hive
Importing RDBMS query results to files (delimited or sequence) in HDFS, or tables in Hive
Exporting files and Hive tables to RDBMS tables

Executes MapReduce jobs to transfer data in parallel with fault tolerance

Download: Microsoft SQL Server Connector for Apache Hadoop from
http://www.microsoft.com/en-us/download/details.aspx?id=27584

Sqoop Import/Export



https://blogs.apache.org/sqoop/entry/apache_sqoop_overview

CloudTech

Marrakesh 2016

Demo: Hive

https://git.io/vrray

CloudTech

Marrakesh 2016

Hands-On Lab

https://git.io/vrraF

Online resources

- Microsoft Azure: azure.microsoft.com
- Free online training at Microsoft Virtual Academy: microsoftvirtualacademy.com
- Lots of stuff: github.com/Azure-Readiness

CloudTech

Marrakesh 2016

Q&A

Francesco Scullino
scullino@ismb.it
v-frscul@microsoft.com

Microsoft | Innovation Center Torino