Hadoop

Jessica Miller

March 17, 2017

Problems with Big Data

Storage

https://www2.cisl.ucar.edu/resources/storage-and-file-systems/hpss

Processing

https://ncar.ucar.edu/community-resources/computational-resources

Jessica Miller

Welcome to Hadoop

Storage

- Divides data
- Stores across multiple nodes

Processing

- Dimension reduction
- Parallel computing

HDFS

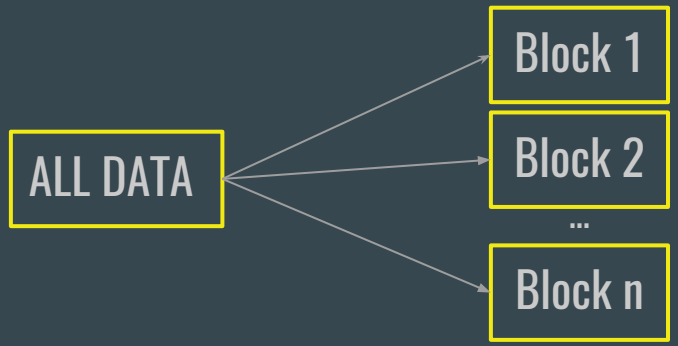
MapReduce

http://www.forbes.com/sites/danwoods/2011/11/03/explaining-hadoop-to-your-ceo/2/#6d30f55d4b54

Jessica Miller

Hadoop Distributed File System

Scalability:



Locality:

Multi-node cluster → common servers → shell + Java

Fail capability:

Block of data → multiple servers

Jessica Miller

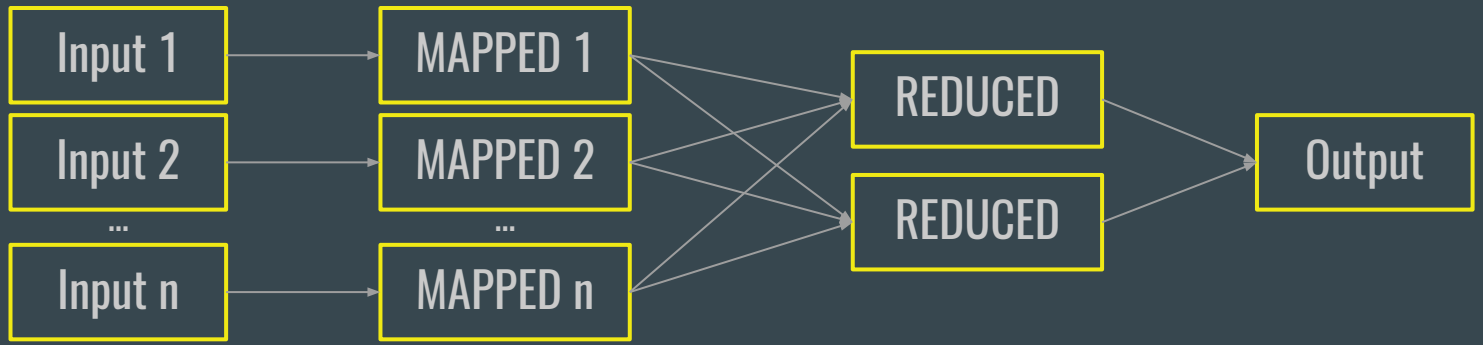
MapReduce

Mapping:

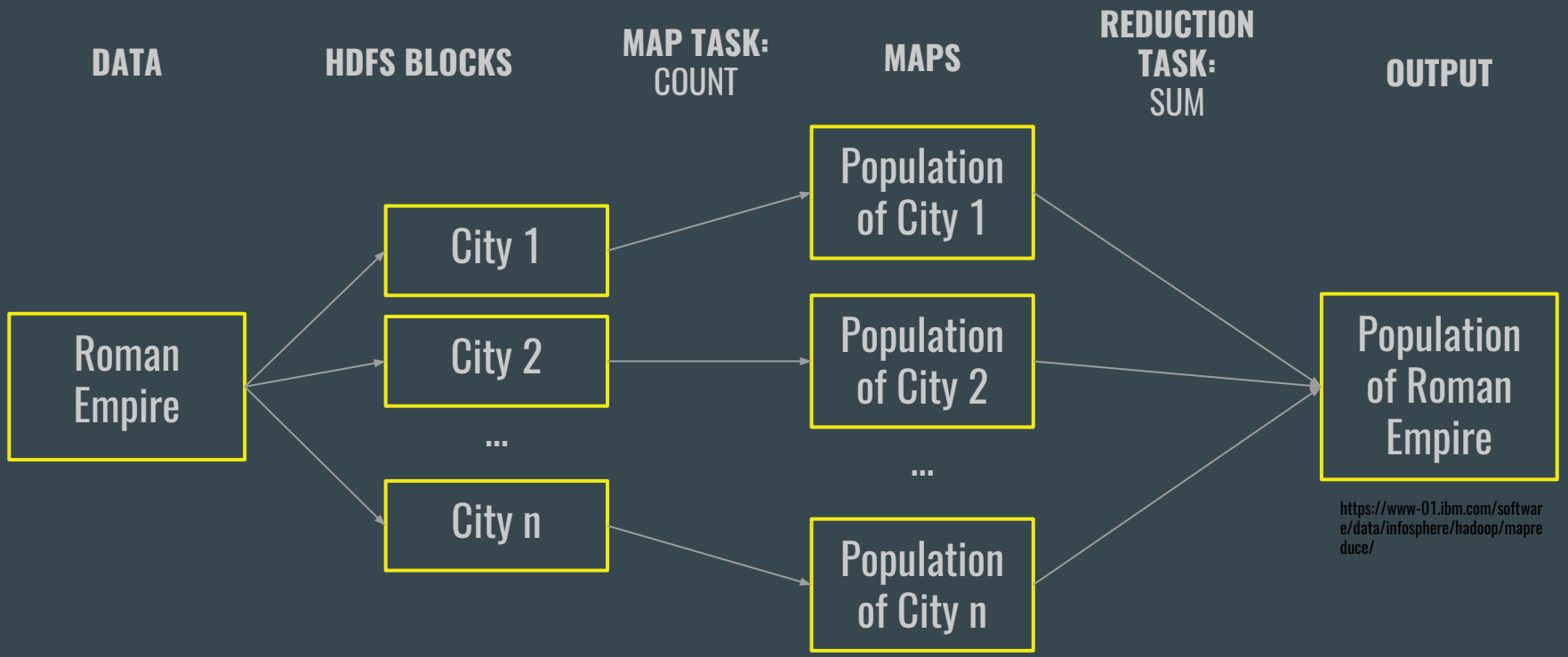
Dataset → mapping task → output

Reduction:

Map output → reduce task → output



Analogy: *Roman Empire Census*



Jessica Miller

Beyond the Fundamentals

- Hadoop YARN : job prioritization, cluster management
- Apache Hive : data querying, summarization, analysis (like SQL)
- Apache Spark: computation over an application
- Apache Pig: parallel computing execution

✓ Open source
✓ Affordable
✓ Projects for easier execution

x Large infrastructure needed
x Straightforward analysis not as easy

Jessica Miller

Who uses Hadoop?

Search/content optimization:

Data storage:

Image conversion:

Jessica Miller

Case Study: *Sears*

Motivation:

- Know the customers better
- Individual customer personalization
- Better retail, higher profit

Results:

- Use of Hadoop framework
- Frontrunner in big data technologies

Before using Hadoop: *Sears*

Gridded data : only 10% analyzed

- Business insights
- Inefficient use of time and money
- Sales not improving

Boldly stepping forward

- Relatively new technology
- Replacement of infrastructure
- Trial and error

http://www.informationweek.com/it-leadership/why-sears-is-going-all-in-on-hadoop/d/d-id/1107038?page_number=1

After using Hadoop: *Sears*

100% of data available for analysis

- 1 node ⟶ 300+ nodes
- Time management
- Money management
- Retail computing and analytics : MetaScale

Individual transactions

https://www.edureka.co/blog/big-data-applications-sears-case-study/

Example: *Sears*

- Query: All items priced more than $29,999.00
- Tools: Ruby MapReduce, Pig, Hive
- Data: 15 billion records
- Solution: Pig can provide quick execution
- Input: 15,274,430,951 records searched
- Output: 28 records returned
- Time: 53 seconds

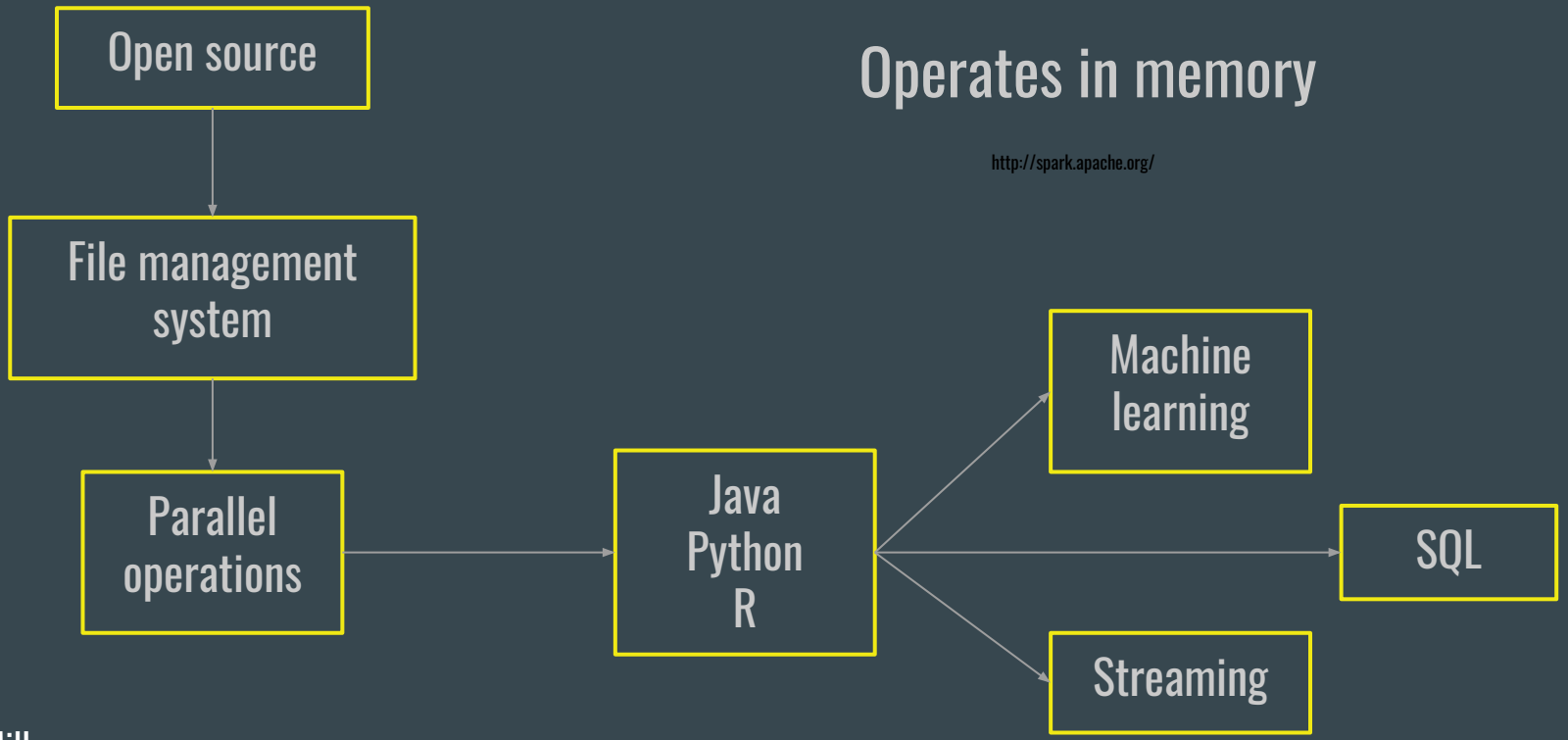
http://www.metascale.com/resources/blogs/156-big-data-case-study-hadoop-first-usage-in-production-at-sears-holdings.html#.WFtg2_krLIU

Further Application: *Apache Spark*

- Data processing specialist
- Up to 100 times faster (in memory) than Hadoop MapReduce
- Built for ease of application usage

http://www.infoworld.com/article/3014440/big-data/five-things-you-need-to-know-about-hadoop-v-apache-spark.html

Step by step: *Spark*



Operates in memory

http://spark.apache.org/

Spark vs. Hadoop

Spark

- Needs data storage system
- In-memory operations
- Reads, operates, writes all at once
- Resilient distributed datasets
- Built-in SQL, machine learning

Hadoop

- Built-in data storage system, HDFS
- Hard drive operations
- Part of the data at a time
- More secure failure capabilities
- Needs advanced analytics add-ons

http://www.forbes.com/sites/bernardmarr/2015/06/22/spark-or-hadoop-which-is-the-best-big-data-framework/#21f1d563532c

From the eyes of a business owner: *Spark*

Do I need advanced analytics?

Is it more cost effective?

Is it easier to execute?

"[...] everyday business owners are finding increasingly innovative uses for their stored data." Bernard Marr, *Forbes*

Thank you!

Google “Hadoop SAS”

https://www.sas.com/en_us/insights/big-data/hadoop.html