Data Warehousing and Data Mining

Lecture 2+

In-depth look at recent trends in data warehousing

Outline

- The shift in paradigm
- Distributed computing in data mining
- MapReduce
- NoSQL

Primer

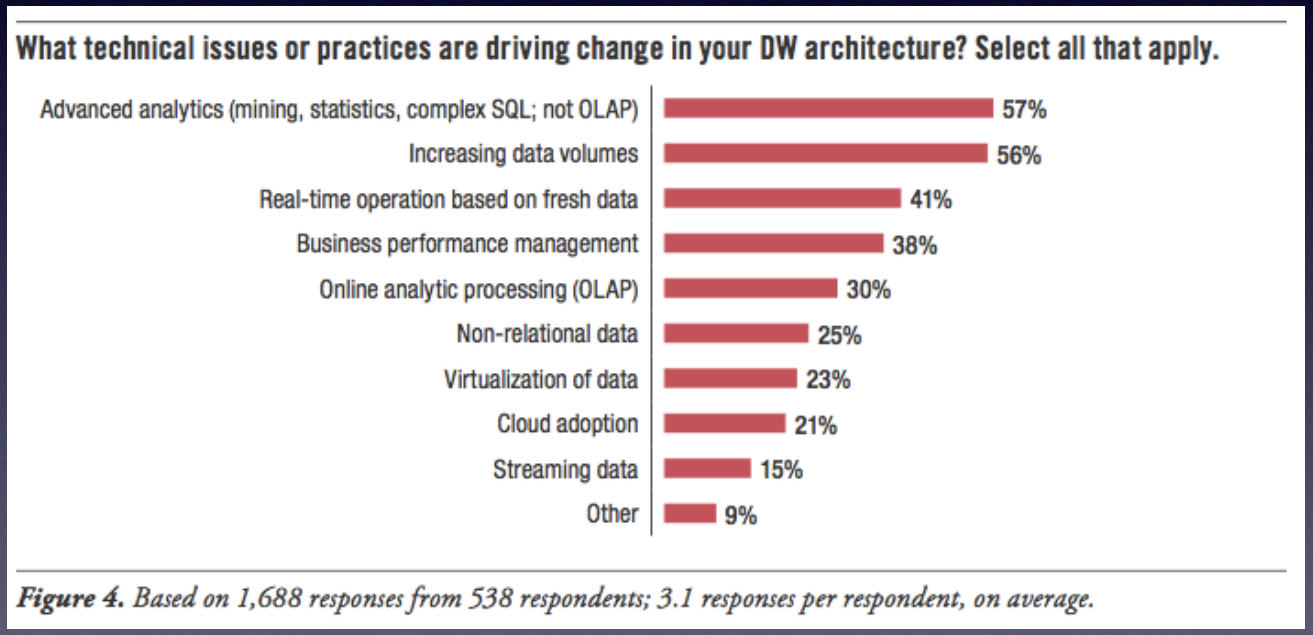
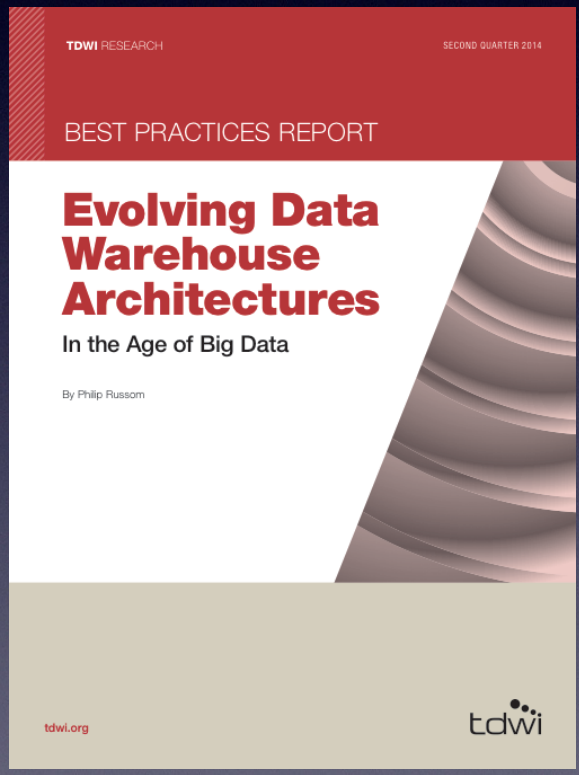


Figure 4. *Based on 1,688 responses from 538 respondents; 3.1 responses per respondent, on average.*



A look at recent trends

- Problem is no longer storage, but analytics
- Data is getting very very big
- Businesses are shifting from simple OLAP to more complex analytics
- Businesses are seeking a more ‘real-time’ solution

Key challenges

- How to **store** data efficiently
- How to **retrieve** data efficiently
- How to **analyse** data efficiently

Solutions

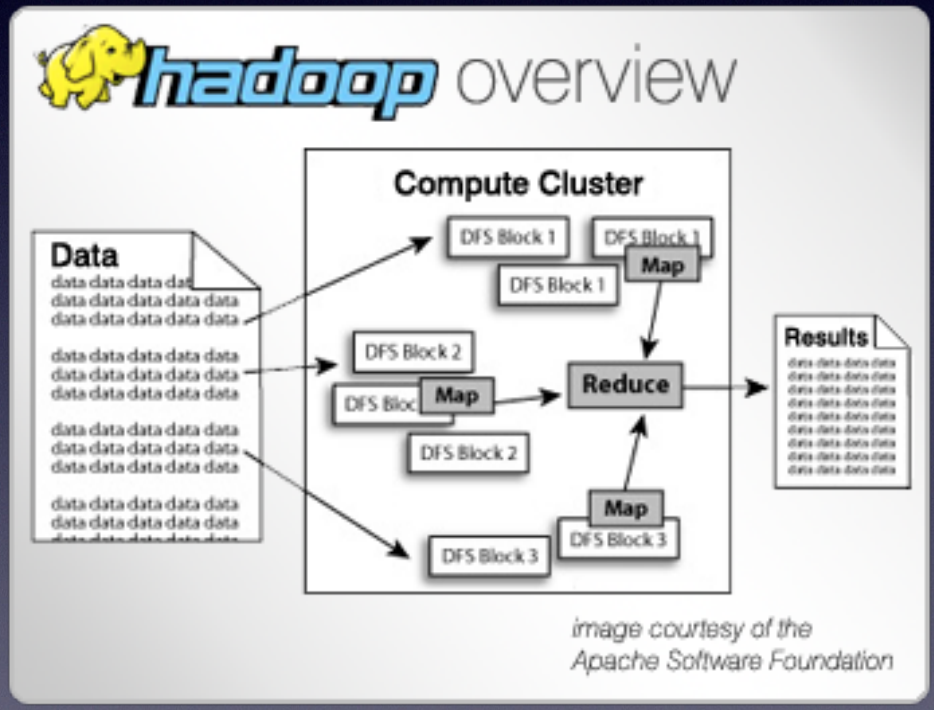
- Scalable architecture
- More than just normal DB storage

Scalable - Hadoop

- Started in 2004
- Aims to be a framework for distributed computing
- Was able to solve many of the key challenges

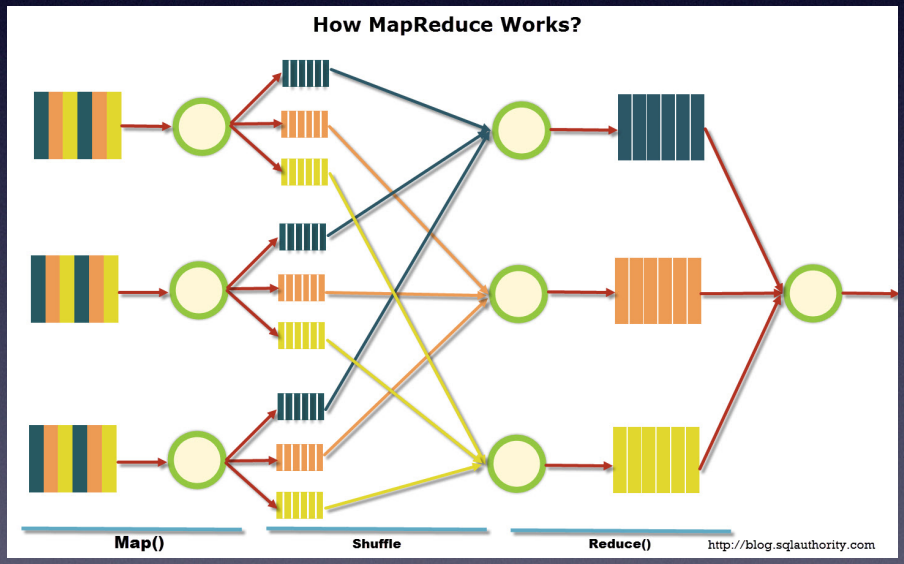
Core modules of Hadoop

- Hadoop Common
- **Hadoop Distributed File System**
- Hadoop YARN
- **Hadoop MapReduce**



MapReduce

- Consist of 2 steps - Map and Reduce
- Map - Split data into small chunks for processing
- Reduce - Aggregate data from many sources



Critique of Hadoop and MapReduce

- Distributed means
  - Concurrency
  - Realtime
  - Horizontally scalable
- However, the downsides are
  - Mapping function must be distributive
  - Often, this limits MapReduce to very simple functions
  - Data must be synced across all nodes that uses it

More than normal SQL - NoSQL

- NoSQL - “Not Only SQL”
- Does not use tabular storage
  - Key-value
  - Object database
  - Graph
  - Document store
- Some people are against it due to the lack of schema

Closing remarks

- MapReduce is making its way to industry standards
- NoSQL is slowly making its way there…some anyways
- Many old-school data scientists don't see NoSQL in the data mining world