3 Big Data Tools

An Introduction

Allan Mitchell
MVP
Copper Blue Consulting

Allan Mitchell

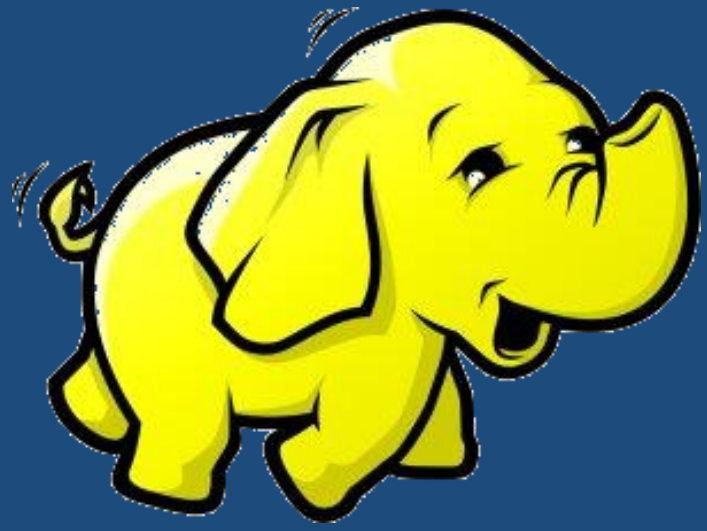
- Joint author on 2005/2008 SSIS Book by Wrox
- Websites
  - www.Copper-Blue.com
- Specialise in Data and Process Integration
- Microsoft SQL Server MVP
- Twitter: allanSQLIS
- E: allan.mitchell@Copper-Blue.com

What is Big Data

- Abused Buzzword/phrase
- TB/PB/ZB
- 3 Vs (OK so there's 5)
  - Volume
  - Velocity
  - Variety
  - (Veracity)
  - (Value)
- When you have to innovate to collect, store, organize, analyse and share it

Why am I talking about Hadoop?

- Hadoop
  - Collaboration with Hortonworks
  - 2 Flavours
    - Azure
    - HDInsight Server



What is Hadoop

- Distributed File System
- Distributed Storage
- Commodity Hardware
- Scales to potentially 1000s of nodes
- Map Reduce
- Developer Loved
- Evolving Ecosystem

What is SQOOP

- SQL to Hadoop
- Actually should be called RELOOP
  - Relational DB to Hadoop
  - Not as catchy though
- Hadoop to SQL
  - SQOOPAndOOPSQ

What is SQOOP

- SQL to Hadoop
  - Files
  - Hive
  - Whole tables, only certain columns, queries
  - Vast array of optional parameters
  - Load particular partitions in Hive
- Hadoop to SQL
  - Only files
  - (technically not true but semantically true)

Why SQOOP

- Good question
  - RDBMS are very good at storing data
  - RDBMS can certainly crunch data
- Let’s look at some potential use cases

Case 1: DW in Hadoop

- Relational databases hold OLTP data
  - Perhaps they hold only a “hot” subset
- Data Warehouse is in Hadoop
  - Need to take relational data and load into Hadoop

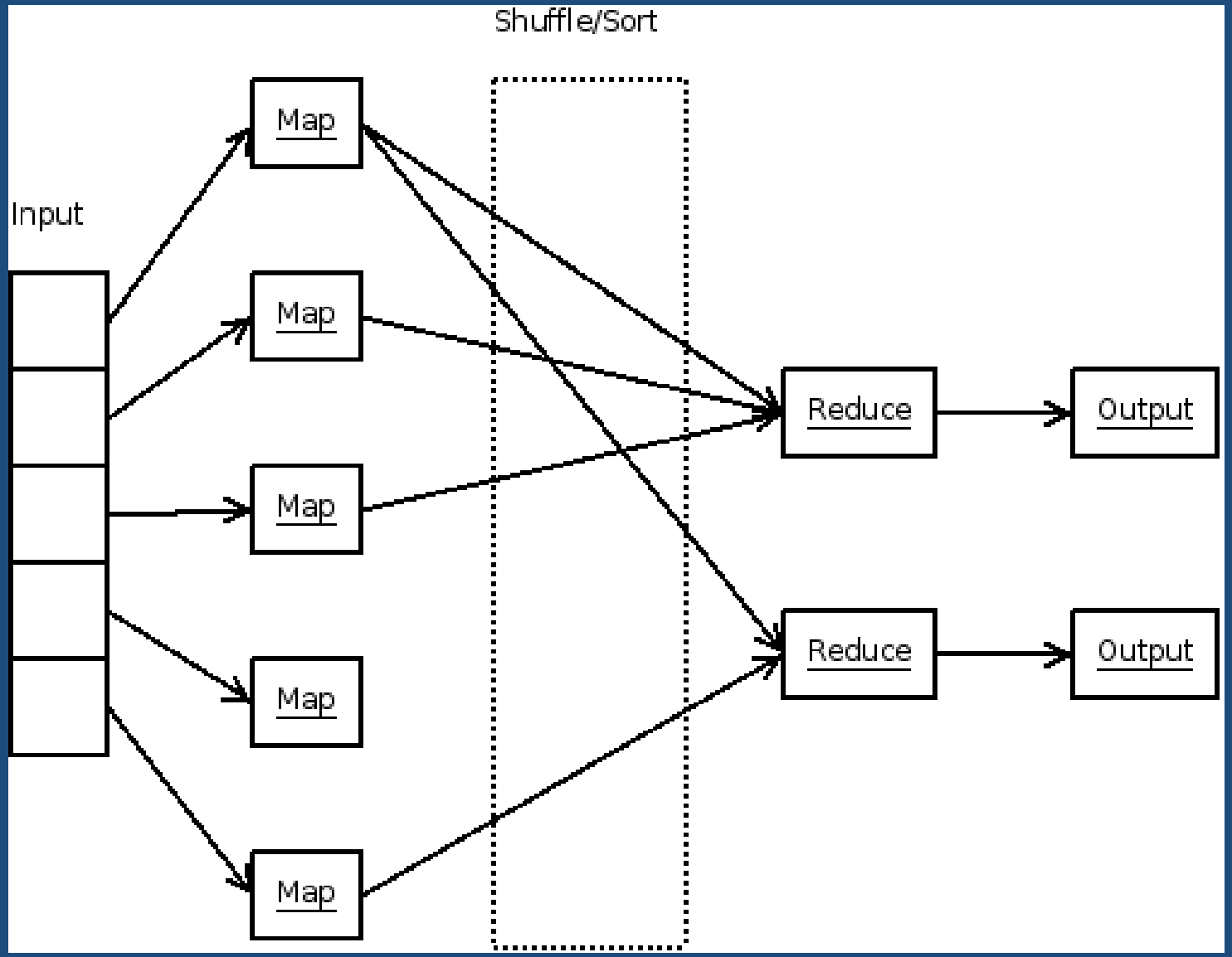
Case 2: Hadoop and MR

- Put data from relational DB into Hadoop
- Use MR to crunch and manipulate data
- Take data from Hadoop back to relational DB

What is Hadoop

- Distributed File System
- Distributed Storage
- Commodity Hardware
- Scales to potentially 1000s of nodes
- Map Reduce
- Developer Loved
- Evolving Ecosystem

What is MR


Shuffle/Sort
Input
Map
Map
Map
Map
Map
Reduce
Reduce
Output
Output

How do I use SQOOP

- Need a JDBC driver
  - Copy to the bin folder of the SQOOP directory
- All command line (No UI)
- Have a lot of patience

Pig

- Hadoop Scripting
- Pig Latin
- Data Flow Expressions
- Efficient
- Can spawn MR jobs
- Grunt
- Schema on read (brilliant)

Pig as a Parallel Dataflow Language

- User describes how data is
  - Read
  - Processed
  - Stored
- SQL describes the answer you want.

For What is Pig Useful?

- ETL
- Raw Data Research/Data Pipelines
- Iterative Programming

How does Hive work?

Hive as a structuring Tool

Creates a schema around the data
Tables stored in Directories

Hive Tables

Rows and columns, like SQL tables

Hive Metastore

Namespace with a set of tables
Holds table definitions
- Physical Layout
- Column Types
- Partition Information

What is the purpose of Hive?

Hive is a data warehousing system for Hadoop

To meet the needs of businesses, data scientists, analysts and BI professionals

Data, Summarized

Fit a structure onto data

Data, Analyzed

Analysis of Large Datasets stored in Hadoop File Systems
SQL-Like language called HiveQL
Custom mappers and reduces when HiveQL isn't enough

HiveQL Queries like SQL Queries?

Similarities in Syntax and Features

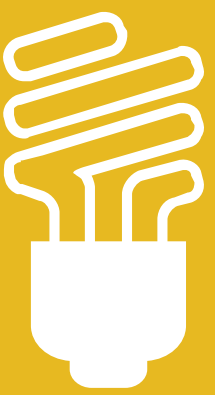
Similar features
SELECT
FROM
WHERE
GROUP BY / HAVING
Table Aliases
Computed Columns

HiveQL Queries like SQL Queries?

Similarities in Syntax and Features

Similar features

Aggregate Functions
Nested Select
CASE
LIKE / RLIKE
JOIN
ORDER BY / SORT BY



Using SQOOP, Pig and Hive

Thank You

allan.mitchell@copper-blue.com
allanSQLIS