Hands-on Introduction to Apache Hadoop and Spark

Douglas Eadline

(Part 2)

BASEMENT SUPERCOMPUTING

Presenter

Douglas Eadline
deadline@basement-supercomputing.com
@thedeadline

- HPC/Hadoop Consultant/Writer
- http://www.basement-supercomputing.com

Outline Day 1

- **Segment 1:** Intro and Overview of Hadoop/Spark (40 mins)
- **Segment 2:** Using HDFS (25 mins)
- Break 10 minutes
- **Segment 3:** Running and Monitoring Hadoop Apps (35 mins)
- **Segment 4:** Using Apache Pig (20 mins)
- Break: 10 minutes
- **Segment 5:** Using Apache Hive (30 mins)

Outline Day 2

- **Segment 6:** Running Apache Spark (35 mins)
- Break: 5 minutes
- **Segment 7:** Running Apache Sqoop (30 mins)
- **Segment 8:** Using Apache Flume (20 mins)
- Break: 10 minutes
- **Segment 9:** Introduction and Example Analytics Application using Apache Zeppelin (40 mins)
- Break: 5 minutes
- **Segment 10:** Wrap-up/ Resources/Where to Go Next (20 mins)

Recommended Approach To Class

- Course covers a lot of material!
- Designed to get you started ("hello.c" approach)
- Sit back and watch the examples
- All examples are provided in a notes file
- I will refer to file throughout the class (cut and paste)
- The notes files are available for download along with some help on installing software (URL provided at end of this class)

User Programming Environment

There are four options (URLs provided at end of course):

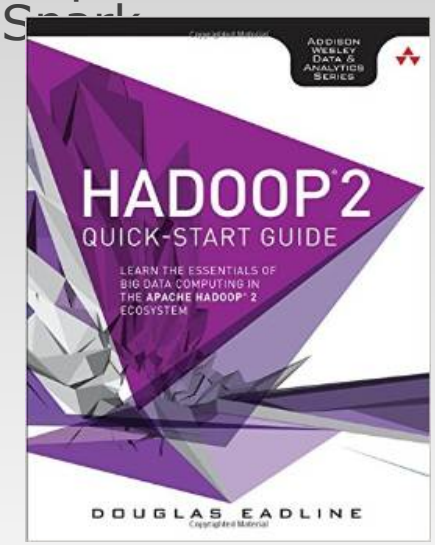
1. Use the Linux Hadoop Minimal Sandbox virtual machine that can run under VirtualBox or VMWare.
2. Use the Hortonworks Sandbox, a full featured Hadoop/Spark virtual machine that runs under Docker, VirtualBox, or VMWare.
3. Install software from Apache.org (instructions provided) Best for Linux machines/laptops
4. Use a resource available to you: local cluster/cloud

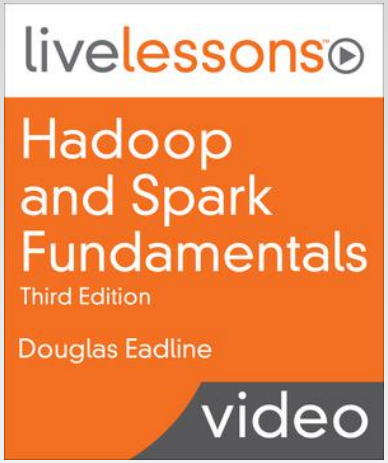
Supporting Materials

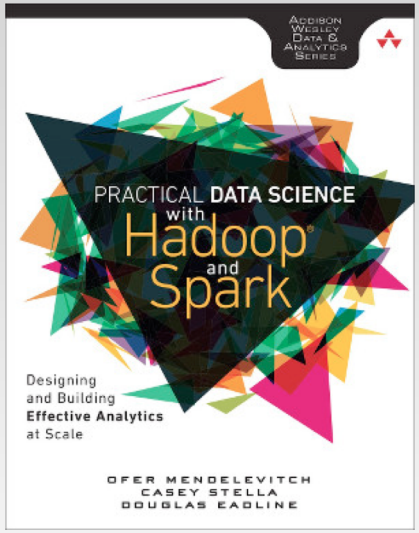
Hadoop 2 Quick-Start: http://www.clustermonkey.net/Hadoop2-Quick-Start-Guide

Hadoop Fundamentals: https://www.safaribooksonline.com/library/view/hadoop-and-spark/9780134770871

Practical Data Science with Hadoop and Spark: http://www.clustermonkey.net/Practical-Data-Science-with-Hadoop-and-Spark







BASEMENT SUPERCOMPUTING

Running Apache Spark

https://spark.apache.org

Apache Spark Background

Spark is a fast and general cluster computing system/language for Big Data. It provides high-level APIs in Scala, Java, Python, and R, and an optimized engine that supports general computation graphs for data analysis. It also supports a rich set of higher-level tools including Spark SQL for SQL and DataFrames, MLlib for machine learning, GraphX for graph processing, Spark Streaming for stream processing, and full integration with Hadoop workflows and HDFS.

Spark RDD and DataFrames

- Resilient Distributed Datasets (RDD) is a fundamental data structure of Spark. DataFrame (DF) is an RDD with named columns (i.e. a database table).
- They are immutable distributed collection of objects.
- Each dataset in RDD/DataFrame is divided into logical partitions, which may be computed on different nodes of the cluster.
- RDDs can contain any type of Python, Java, R, or Scala objects, including user-defined classes. DataFrames work like database tables.

Spark Operations: Transformations and Actions

- **RDD/DF Transformations** return a pointer to new RDD/DF . The original RDD/DF cannot be changed. Spark is lazy, so nothing will be executed unless unless results of a transformation or an action is called.

  An RDD transformation is not a set of data, but is a step in a program (might be the only step) telling Spark how to get data and what to do with it.
- **RDD Actions** return values (e.g. collect, count, take, save-as).

5 Minute BREAK

Segment 7

Running Apache Sqoop

http://sqoop.apache.org

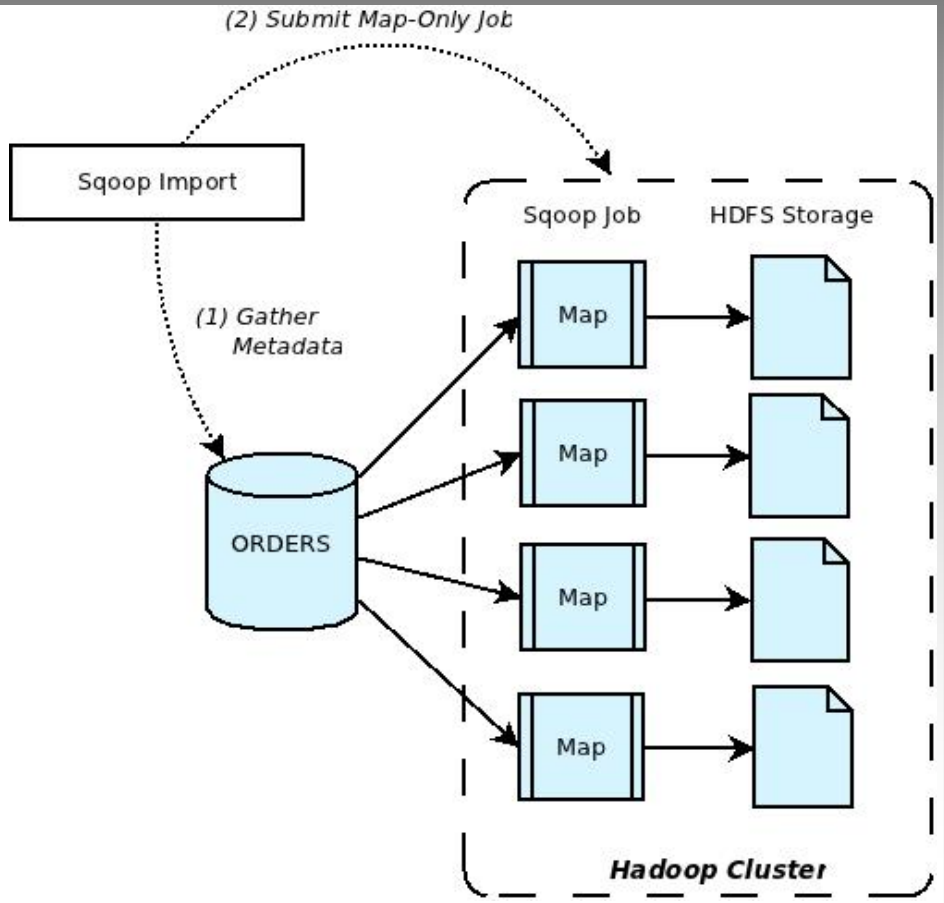
BASEMENT SUPERCOMPUTING

Apache Sqoop Background

Sqoop is a tool designed to transfer data between Hadoop and relational databases or mainframes. You can use Sqoop to import data from a relational database management system (RDBMS) such as MySQL or Oracle or a mainframe into the Hadoop Distributed File System (HDFS), transform the data in Hadoop MapReduce, and then export the data back into an RDBMS.

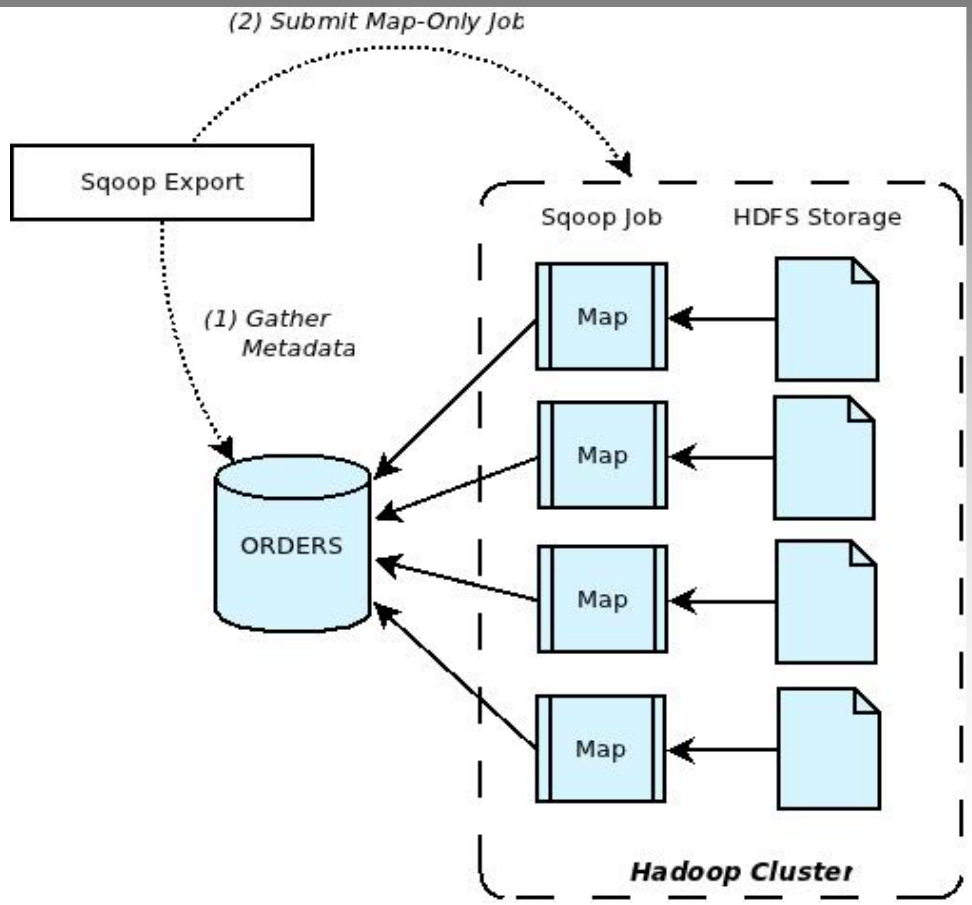
Sqoop Data Import

- Sqoop works with local RDBMS
- Sqoop gathers metadata about DB
- Can work sequentially or in parallel
- In parallel each map process contacts the RDBMS system and "pulls" it's slice of data



Sqoop Data Export

- Sqoop works with local RDBMS
- Sqoop gathers metadata about DB
- Can work sequentially or in parallel
- In parallel each map process contacts the RDBMS system and "pushes" it's slice of data



Segment 8

Using Apache Flume

https://flume.apache.org

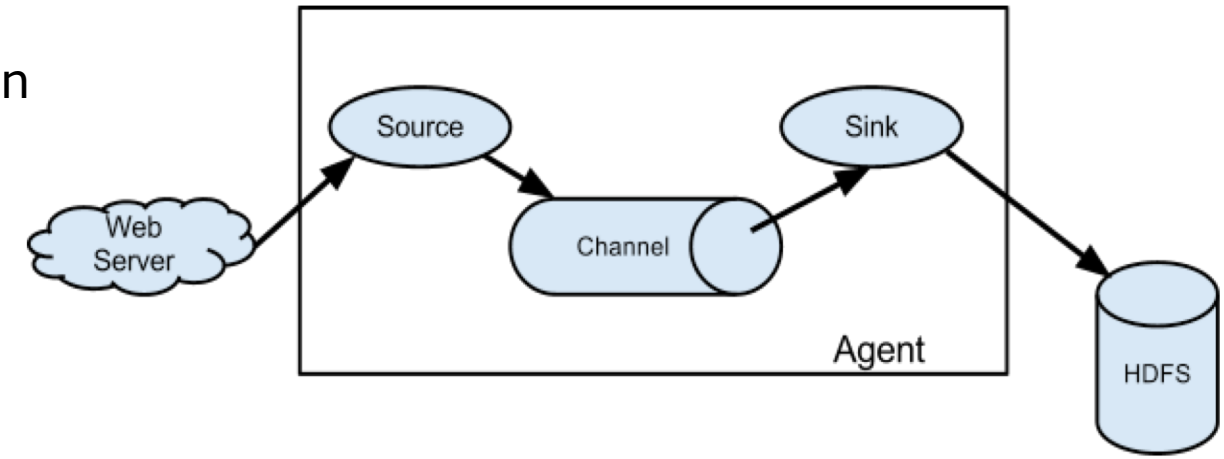
BASEMENT SUPERCOMPUTING

Apache Flume Background

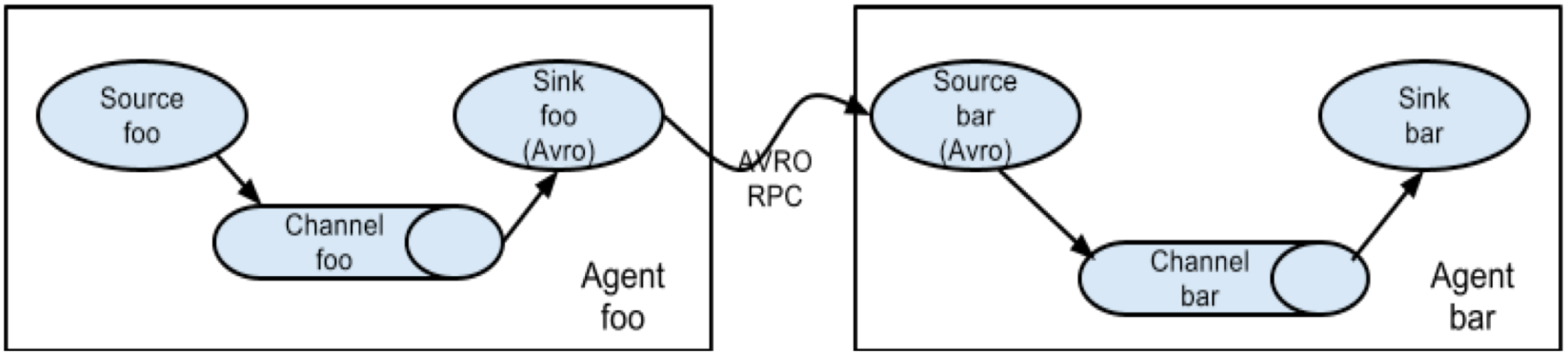
Apache Flume is a distributed, reliable, and available service for efficiently collecting, aggregating, and moving large amounts of streaming data into the Hadoop Distributed File System (HDFS). It has a simple and flexible architecture based on streaming data flows; and is robust and fault tolerant with tunable reliability mechanisms for failover and recovery.

Flume Agent

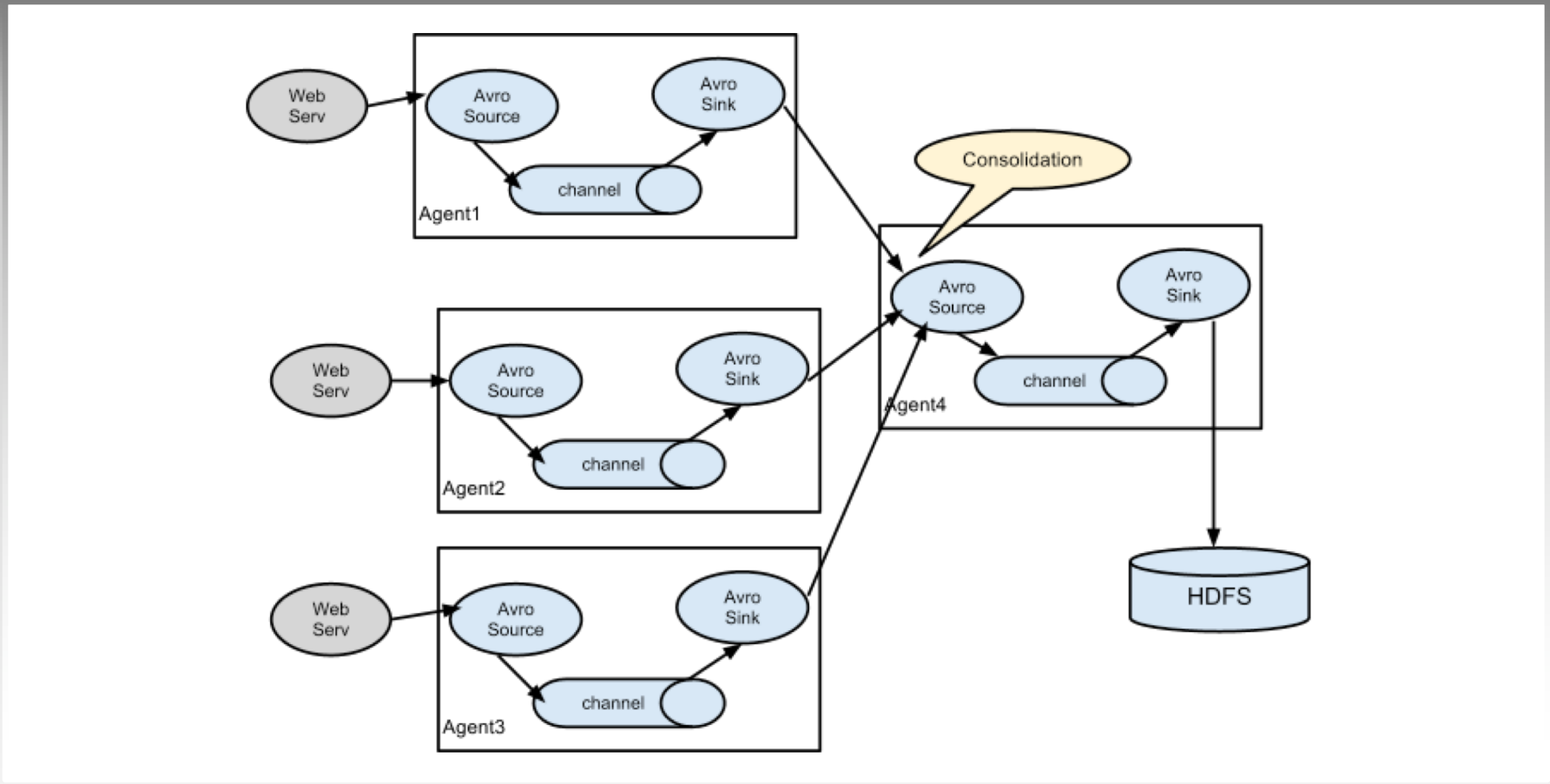
- A **Source** receives data and sends it to a channel. It can send the data to more than one channel.
- A **Channel** is a data queue that forwards/buffers source data to the sink.
- A **Sink** delivers data to destination such as HDFS, a local file, or another Flume agent.



Flume Pipeline Multiple Agents



Flume Consolidate Multiple Sources



5 Minute BREAK

Segment 9

A Walking Tour of the Apache Zeppelin Web

https://zeppelin.apache.org

BASEMENT SUPERCOMPUTING

Apache Zeppelin Background

- Web-based notebook that enables data-driven, interactive data analytics and collaborative documents with SQL, Scala and more.
- Apache Zeppelin interpreter concept allows any language/data-processing-backend to be plugged into Zeppelin. Currently Apache Zeppelin supports many interpreters such as Apache Spark, Hive, SparkR, PySpark, JDBC, Markdown, and Shell.

Segment 10

Example Analytics Application with Apache Zeppelin

https://zeppelin.apache.org

BASEMENT SUPERCOMPUTING

5 Minute BREAK

Segment 11

Wrap-up
Resources
Where to Go Next

Course Take Aways

- Almost all tools have good on-line documentation and further examples
- The course examples will help you get started using these tools (worked examples)
- The tool to use depends on what you want to do!
- Spark is a good place to start because it has all the needed components.
- Hadoop is a platform on which to build your analytics applications

Course Resources

Download all Class Notes and data files used in the lessons. Includes directions on how to install Hadoop Core, Pig, Hive, Spark, Sqoop, Flume, and Zeppelin on a Linux host.

https://www.clustermonkey.net/download/Hands-on_Hadoop_Spark

BASEMENT SUPERCOMPUTING

Hadoop/Spark/Zeppelin Resources

Hadoop-Minimal Sandbox VM

(Supports all tools used in this course, including Zeppelin)

- http://www.clustermonkey.net/download/Hands-on_Hadoop_Spark

  Linux-Hadoop-Minimal.ova

  Linux-Hadoop-Minimal-Install.txt

- At a minimum 2 cores, 4 GB RAM, 70G disk space

Hortonworks Sandbox VM

- https://hortonworks.com/products/sandbox
- At a minimum 4 cores, 16 GB RAM, 70G disk space

BASEMENT SUPERCOMPUTING

Questions ?

Thank you