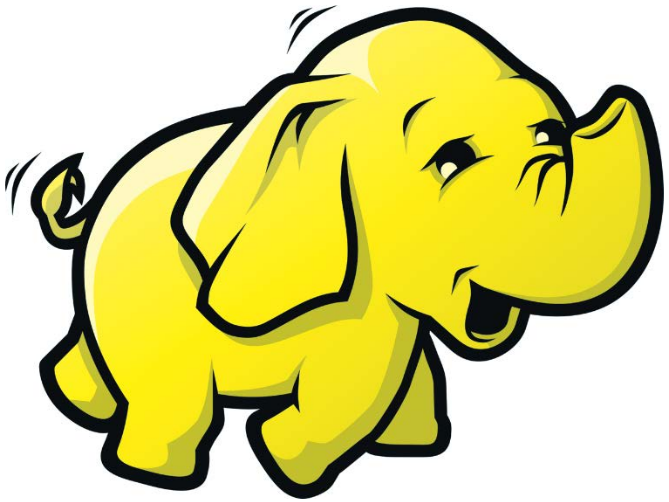
Automating Basic HDFS Operations

Thomas M. Henson

@henson_tm | www.thomashenson.com

Overview



Bash scripts for ingesting

Why automate ingest into HDFS

Walk though a real life demo

Next steps

Bash

Unix scripting language that processes POSIX commands. Widely used by System Administrators to automate routine tasks in POSIX environments.

Bash Scripting

Runs on Cluster

MapReduce Jobs

Easy to Learn

Example Bash Script

```
#!/bin/bash
files=“1 2 3 4 5"

for fileName in $files
do
  echo "$fileName"
done
```

Ingesting Multiple Files

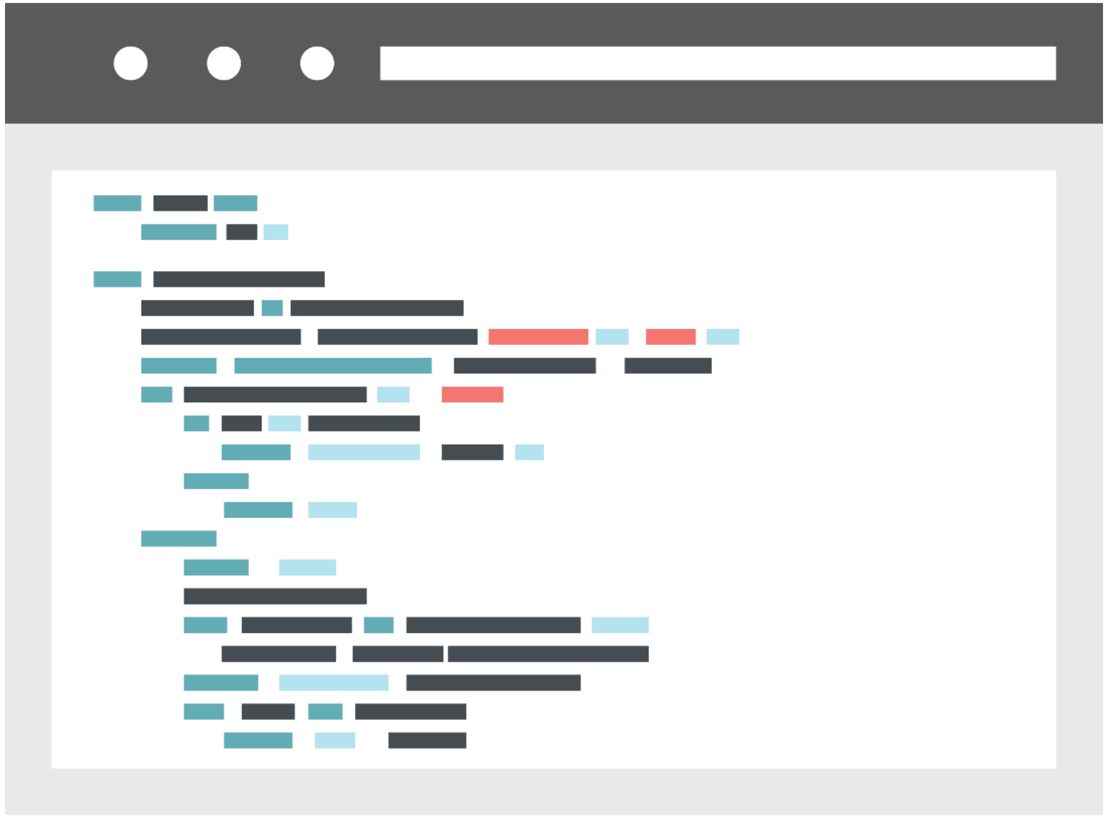
- Using command line for single files
  - HDFS dfs –put /tmp/somefile /user/somefile
- Multiple files
  - 10
  - 100
  - 1 Million
- HDFS built to scale

Speed up Ingesting

- Lazy developer says, “Automate any redundant task possible.”
- Use our knowledge of HDFS to automate ingestion
- Use our Linux knowledge

Data Ingest in HDFS

- HDFS DFS
- Java
- Pig
- Sqoop
- Python
- Bash



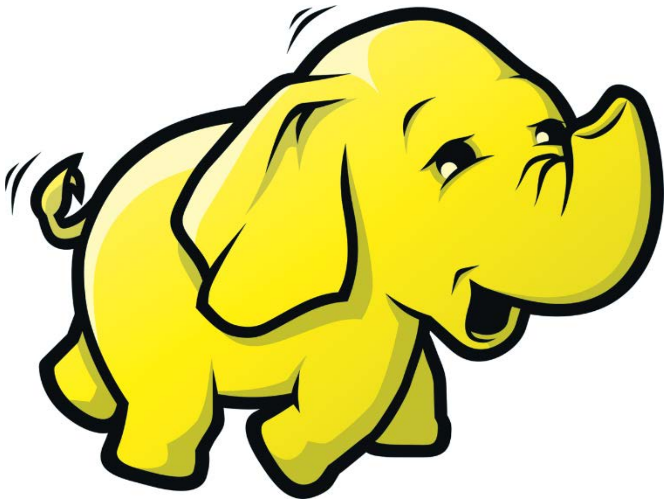
Demo

Simple Bash Script

Demo 2

Walk through ingesting multiple files

Summary



Understand why we should automate

Talked about ingesting data in HDFS

Looked at Bash script for HDFS

Hands on Bash scripts

Congratulations

HDFS from the Command

Keep Learning

- Documentation is your Friend
- Answer Questions on Discussion Post
- Pluralsight Courses
  - Hive
  - Pig
  - Scala
- Contribute to an open source project