Building a Data Pipeline

with Tools from the Apache Hadoop Ecosystem

Rich Haase
Twitter - @richhaase
LinkedIn - linkedin.com/in/richhaase

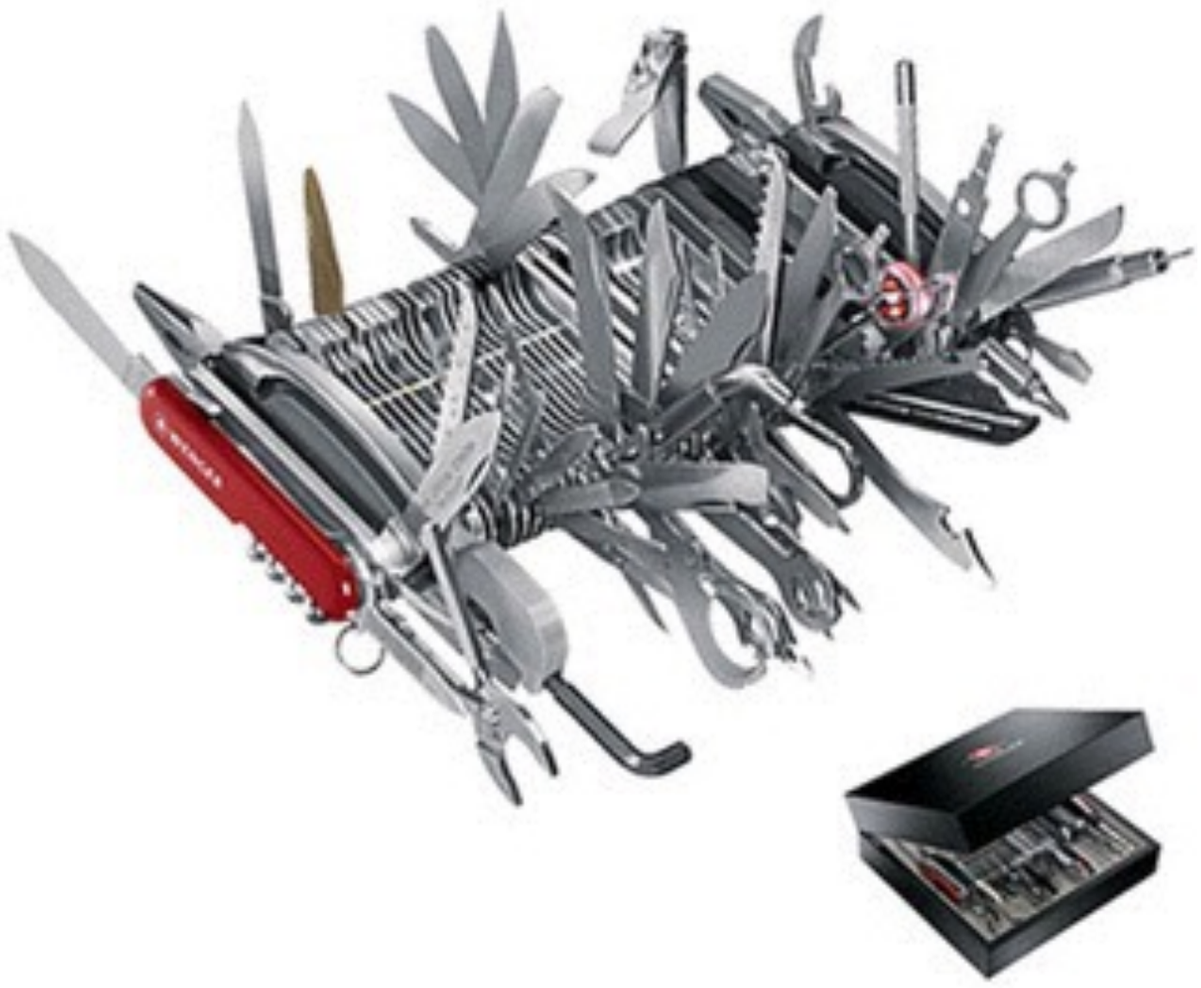
About Me

- 18 years experience in technology
- Infrastructure/Data
- Started using Hadoop in 2010 and haven't looked back
- *Mildly* obsessed with the movement and management of lots of data

Why Hadoop?

This is your home grown data pipeline toolkit.

http://www.heavycoverinc.com/heavy-cover-titanium-spork-multi-tool/



This is your data pipeline toolkit when you make use of the Hadoop ecosystem.

http://www.overstock.com/Sports-Toys/Wenger-Giant-85-tool-141-function-Swiss-Army-Knife/3361457/product.html

Distribution Software Matrix

	CDH 5.7.1	HDP 2.4.2	Bigtop 1.1	MapR 5.1	EMR 4.7.0
Flume	1.6.0	1.5.2	1.6.0	1.6.0	-
Hadoop	2.6.0	2.7.1	2.7.1	2.7.1*	2.7.2
Hue	3.9.0	2.6.1	3.9.0	3.9.0	3.7.1
HBase	1.2.0	1.1.2	0.98.12	1.1	1.2.1
Hive	1.1.0	1.2.1	1.2.1	1.2.1	1.0.0
Oozie	4.1.0	4.2.0	4.2.0	4.2.0	4.2.0
Pig	0.12.0	0.15.0	0.15.0	0.15.0	0.14.0
Spark	1.6.0	1.6.1	1.5.1	1.6.1	1.6.1
Sqoop	1.4.6	1.4.6	1.4.5	1.4.6	1.4.6
ZooKeeper	3.4.5	3.4.6	3.4.6	-	3.4.8

https://docs.aws.amazon.com/ElasticMapReduce/latest/ReleaseGuide/emr-release-components.html
https://www.cloudera.com/documentation/enterprise/release-notes/topics/cdh_vd_cdh_package_tarball_57.html
http://www.apache.org/dist/bigtop/bigtop-1.1.0/repos
http://docs.hortonworks.com/HDPDocuments/HDP2/HDP-2.4.2/bk_HDP_RelNotes/content/ch_relnotes_v242.html
http://maprdocs.mapr.com/home/#InteropMatrix/r_eco_matrix.html

The Hadoop Ecosystem

- ~150 projects
- Dozens of vendors
- Contributions from a wide variety of organizations and individuals
- Constantly evolving

https://hadoopecosystemtable.github.io/

Tier 1

Included in all major distributions

Tier 2

Included in a major distribution

Apache DataFu™

Parquet

CASCADING

AVRO

APACHE STORM™

accumulo

Apache orc

APACHE DRILL

APACHE PHOENIX

mahout

Sentry

CRUNCH

presto

TEZ

APACHE FALCON

APACHE KNOX

Apache Atlas

Apache Ranger

Apache Ambari

KUDU

apache myriad

Tier 3

Not included in any distribution

Apache Ignite
Flink
Apache Twill
APEX
HAWQ
TAJO
REEF
Apache Trafodion
OpenTSDB
nifi
chukwa
HELIX
ML
CDAP
CASK DATA APPLICATION PLATFORM
APACHE GIRAPH
buildoop
talend
druid
Luigi
PigPen
kiji
samza

Disclaimer:

I’m sure I’ve missed projects on this list. Any oversight was completely unintentional.

Tiers are not indicative of software quality, nor is it an indictment of the engineers/organizations who contributed to the project. Every open source contribution brings a fairy back to life.

Also, some projects didn’t have logos.


www.whatsmeme.com
NOW WHAT

Data Life Cycle

Capture

Enrichment

Analysis

Presentation

Reporting

Archival

Removal

Demo

https://github.com/richhaase/building-a-data-pipeline

Questions

See https://github.com/richhaase/building-a-data-pipeline/blob/master/README.md for references.