cloudera
Ask Bigger Questions

How to Build a Hadoop Data Application

Tom White @tom_e_white

Strata London #strataconf

13 November 2013

Agenda

- Logistics
- Introduction to Hadoop and CDK
- Tutorial
- Wrap Up

Logistics

Laptop set up

- VirtualBox (preferred), VMware Player, or VMware Fusion
- Cloudera QuickStart VM
  - Install now from USB memory stick if you haven't already
  - https://github.com/cloudera/cdk-examples/tree/0.7.0/demo#troubleshooting
- Install CDK examples

```
cd cdk-examples
git pull origin 0.7.0
cd dataset
mvn install
```

Administration with Cloudera Manager

- Login: admin/admin
- Stop and start services
- Update configuration
  - Click on “Configuration” tab. Click “View and Edit”
  - Restart service

Hue: the Hadoop Web UI

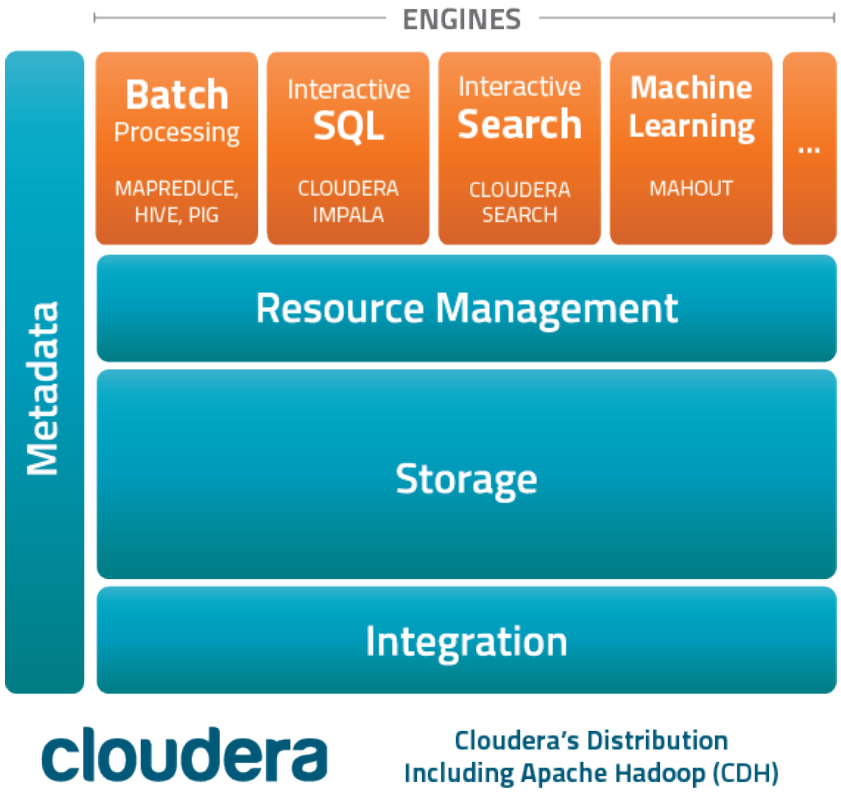
- Login: cloudera/cloudera
- Useful services
  - Beeswax (Hive UI)
  - Impala Query UI
  - File Browser

Introduction to Hadoop and CDK

What is Hadoop?

HDFS and MapReduce

A Hadoop Stack



cloudera

Cloudera's Distribution
Including Apache Hadoop (CDH)

Glossary

- Apache Avro – cross-language data serialization library
- Apache HCatalog – metadata storage system (part of Hive)
- Apache Flume – streaming log capture and delivery system
- Apache Oozie – workflow scheduler system
- Apache Crunch – Java API for writing data pipelines
- Parquet – column-oriented storage format for nested data
- Impala – interactive SQL on Hadoop

Hadoop Pain Points*

* Not exhaustive

Choosing a File Format

	No compression	gzip	snappy	lzo	bzip2
Delimited text					
JSON					
Sequence File		?	?	?	
Avro File					
RCFile					
Parquet					
…					

Defining a Data Model

Schema on read
vs.
Schema on write

What is the user ID field called?

- uid
- userId
- userid
- user_id
- user_Id

- “Scaling Big Data Mining Infrastructure: The Twitter Experience” by Lin and Ryaboy

Bill Graham
@billgraham

Follow

Ah! I just had my first run in with the dreaded camel_Snake in our code base. Whoa, that was scary.

Reply Retweet Favorite More

3 RETWEETS | **1** FAVORITE

5:58 PM - 11 Sep 12

Defining a File Layout

- Which is best?
  - /data/clickstream/20120101
  - /data/clickstream/date=20120101
  - /data/clickstream/2012/01/01
  - /data/clickstream/year=2012/month=01/day=01

A Pattern: Hadoop is Flexible

… but also low-level and complex

Some Best Practices

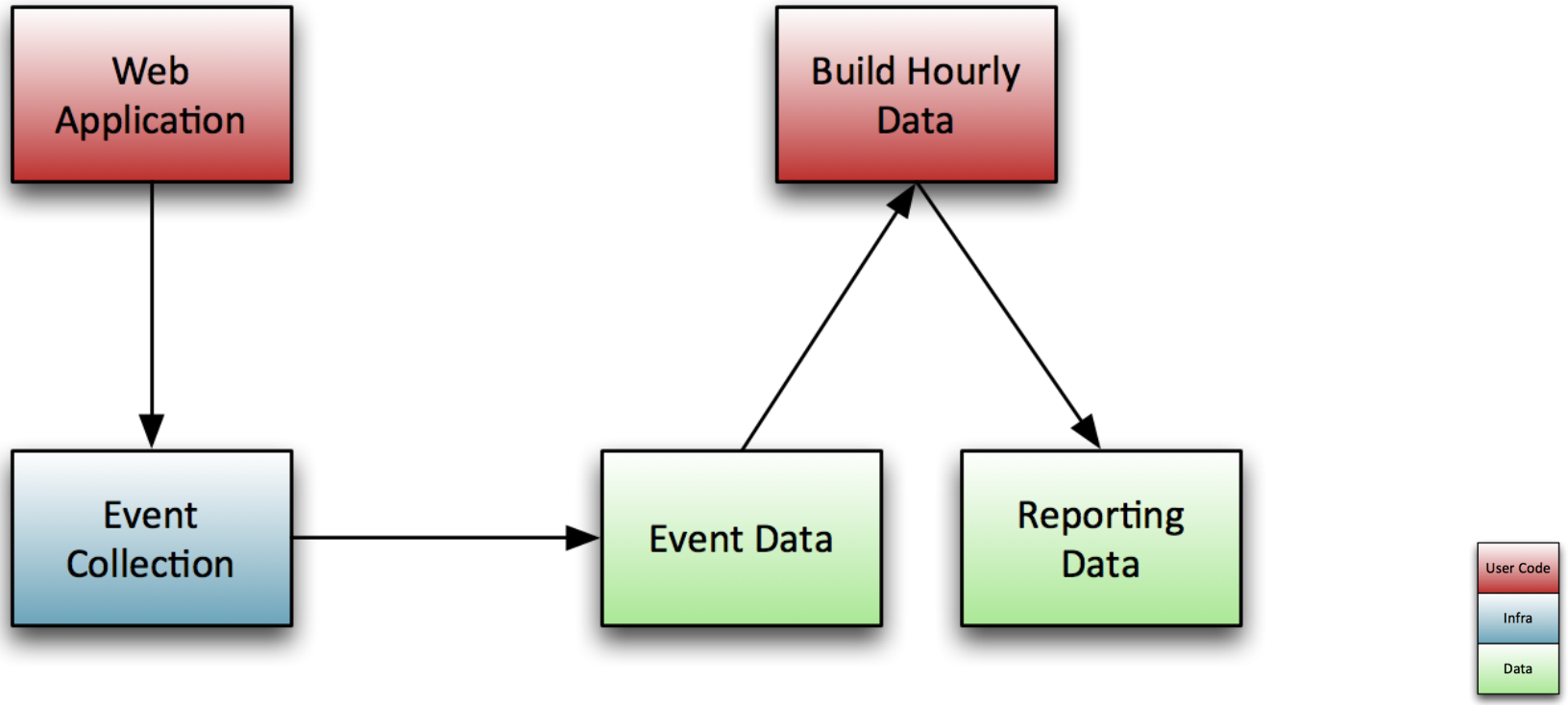
- Use Avro Schemas for the data model
- Use Avro Data Files for row-oriented data
- Use Parquet for column-oriented data
- Use a Hive/HCatalog compatible file layout:
  - /data/<dataset>/partition-1=<x>/partition-2=<y>
- Use a library like Crunch or Cascading for batch analysis
- Use Impala for interactive ad hoc analysis

The Cloudera Development Kit Codifies Best Practice as APIs, Tools, Docs and Examples

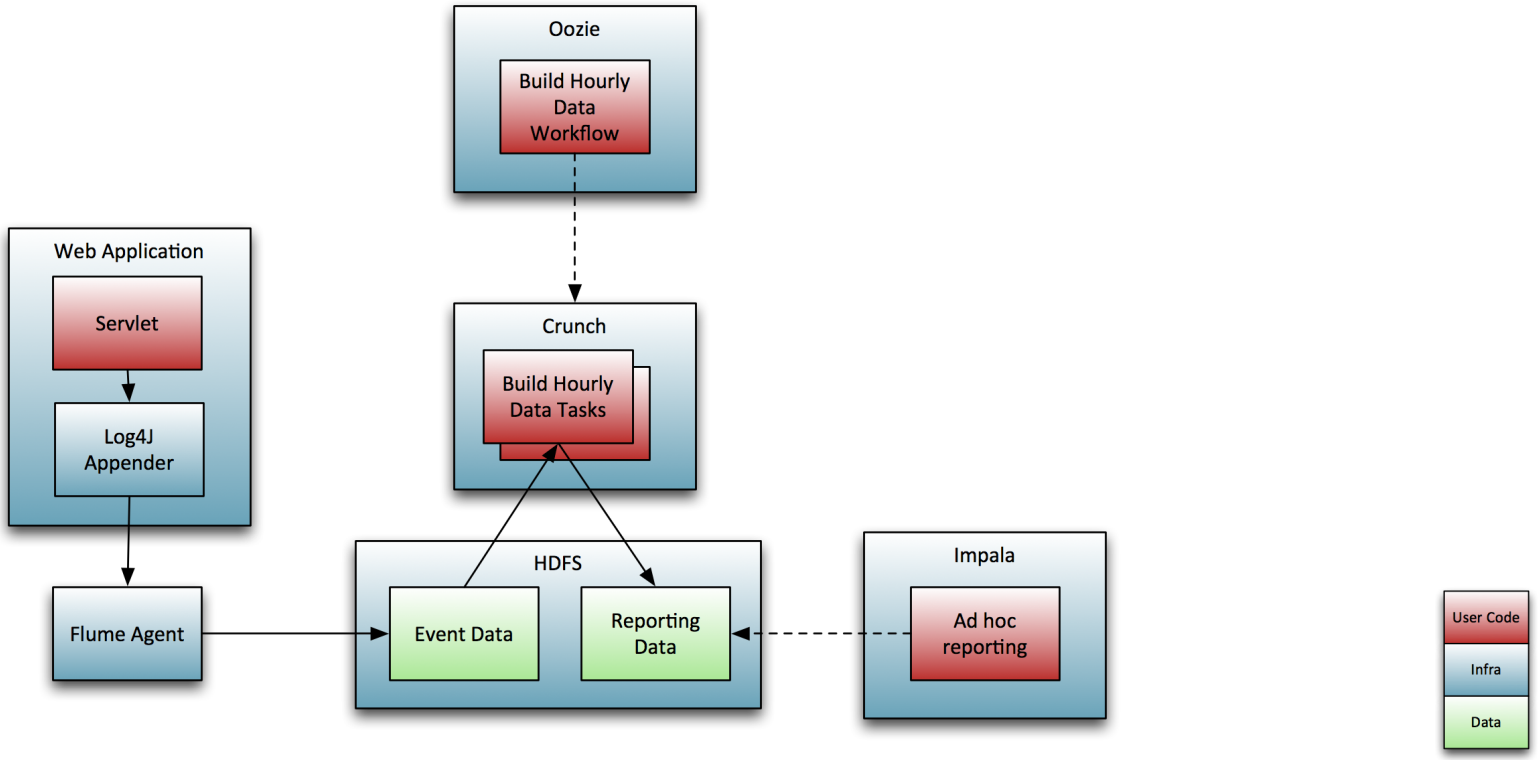
CDK

- A client-side library for writing Hadoop Data Applications
- First release was in April
- 0.8.1 released last month
- Open source, Apache 2 license
- Modular
  - Data module (HDFS, Flume, Crunch, HCatalog)
  - Morphlines transformation module
  - Maven plugin

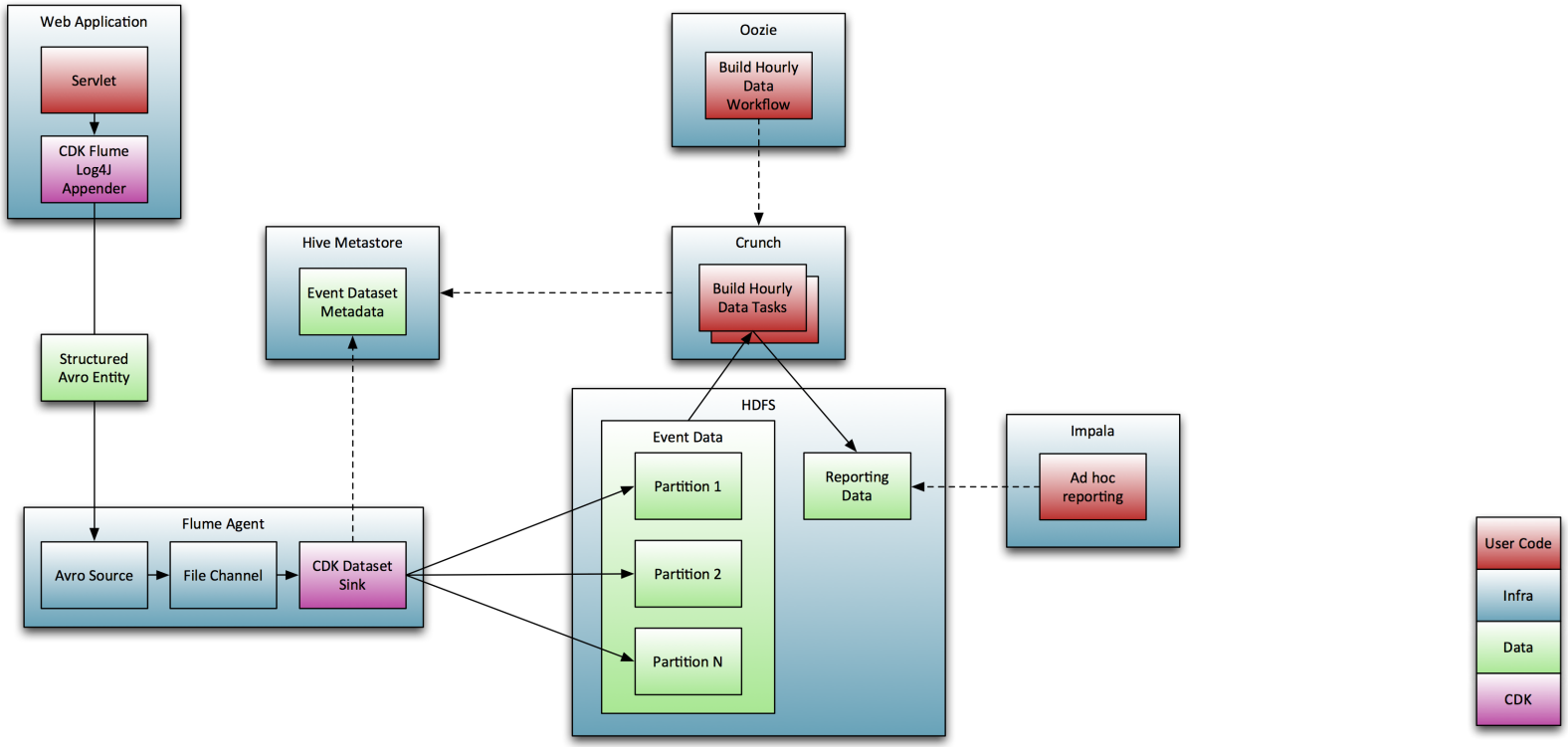
A typical system (zoom 100:1)



A typical system (zoom 10:1)



A typical system (zoom 5:1)



Tutorial

Dataset example

- In this example you will:
  - Create a Products dataset in HDFS
  - Write records to the dataset
  - Read records from the dataset
  - Drop the dataset
- Follow the instructions at
  - https://github.com/cloudera/cdk-examples/tree/0.7.0/dataset

Dataset API

- Dataset – a collection of entities
- DatasetRepository - physical storage location for datasets
- DatasetDescriptor – holds dataset metadata (schema, format)
- DatasetWriter – write entities to a dataset in a stream
- DatasetReader – read entities from a dataset in a stream

End-to-end data pipeline example

- In this example you will:
  - Log application events from a webapp using Flume
  - Extract session data from the events using Crunch
  - Run the Crunch job periodically using Oozie
  - Analyze session data with SQL using Impala or Hive
- Follow the instructions at
  - https://github.com/cloudera/cdk-examples/tree/0.7.0/demo
  - Note: skip “Configuring the VM” section

Wrap Up

Ideas for what to try next

- Run on a real cluster
- Add more derived datasets
- Customize the data model
- Use Morphlines for data transformations

CDK Resources

- Docs
  - http://cloudera.github.io/cdk/docs/current/
- Examples
  - https://github.com/cloudera/cdk-examples
- Mailing list
  - https://groups.google.com/a/cloudera.org/forum/#!forum/cdk-dev


cloudera®
Ask Bigger Questions