cloudera
Ask Bigger Questions

Real World Hadoop Use Cases

JFokus 2013, Stockholm

Eva Andreasson, Cloudera Inc.

Lars Sjödin, King.com

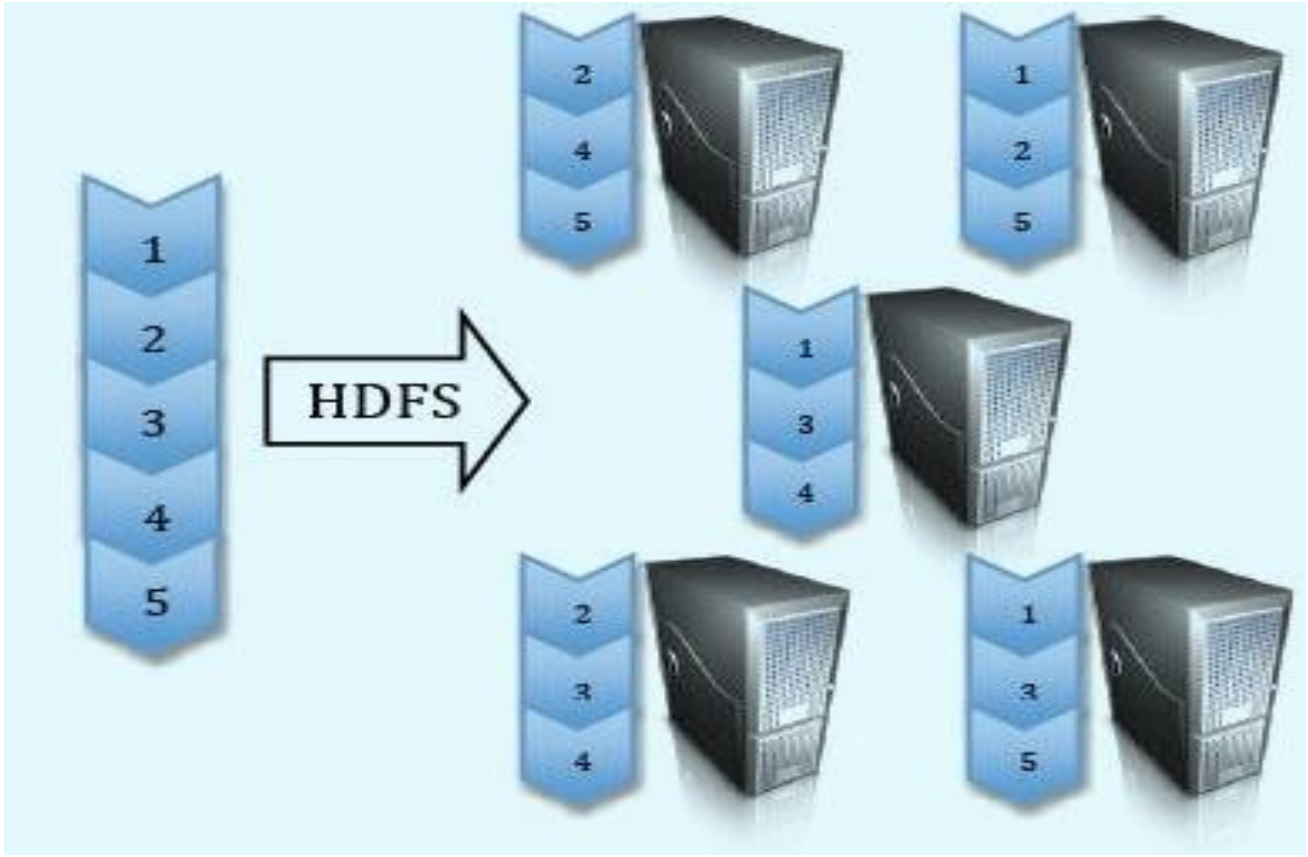
Agenda

- Recap of Big Data and Hadoop
- Analyzing Twitter feeds with Hadoop
- Real world Hadoop use case Featuring King.com
- Q&A

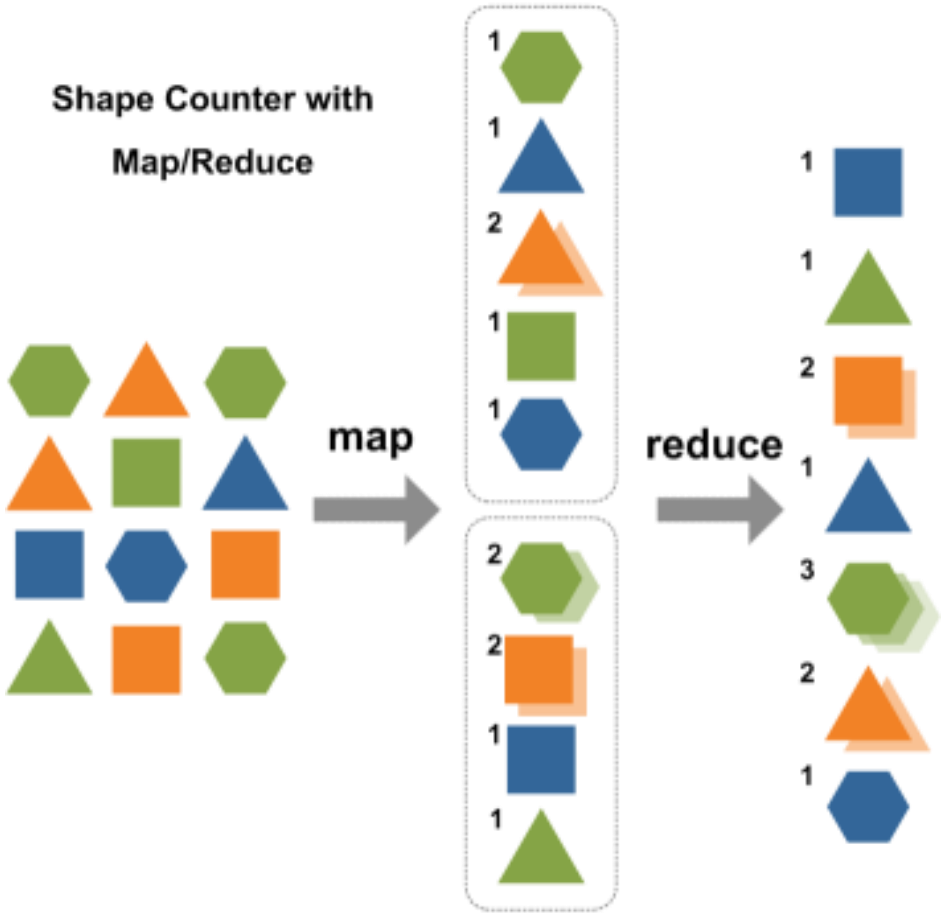
Big Data?

- Big Data
  - Increased volumes of data
  - Increased speed of incoming data
  - Increased variety of data types
- Challenges
  - Stress on traditional systems
    - Process more data within same time window
    - ETL / Cleansing of exponential ingest amounts and new data types
  - Inflexible models for when questions change
  - Siloed data / organizations preventing “most value”

Hadoop Distributed File System (HDFS)



MapReduce: A scalable data processing framework



REAL WORLD EXAMPLE #1

ANALYZING TWITTER DATA WITH HADOOP

Analyzing Twitter

Hint: @EvaAndreasson

- Social media popular with marketing teams
- Twitter is an effective tool for promotion

- But how do we find out who is most influential:
  - Who is influential and has the most followers?
  - Which Twitter user gets the most retweets?
  - Who is influential in our industry?

Techniques

- SQL
  - Filter on industry
  - Aggregate tweets by original poster and count retweets
  - Sort
- Complex data
  - Deeply nested
  - Variable schema
  - Size of data set

Hadoop!

Flume

- Streaming data flow (like Twitter)
- Sources
  - Push or pull
- Sinks
- Event based

Pulling data From Twitter

- Custom source, using twitter4j
- Source will process data as discrete events
  - Filter on key words
- Sink writes to files in HDFS

Loading data into HDFS

- HDFS Sink comes stock with Flume
- Easily separate files by creation time
  - `hdfs://hadoop1:8020/user/flume/tweets/%Y/%m/%d/%H/`

Outline of Flume Source for Tweets

```
public class TwitterSource extends AbstractSource
    implements EventDrivenSource, Configurable {
  ...
  // The initialization method for the Source. The context contains all
  // the Flume configuration info
  @Override
  public void configure(Context context) {
    ...
  }
  ...
  // Start processing events. Uses the Twitter Streaming API to sample
  // Twitter, and process tweets.
  @Override
  public void start() {
    ...
  }
  ...
  // Stops Source's event processing and shuts down the Twitter stream.
  @Override
  public void stop() {
    ...
  }
}
```

Twitter API

- Callback mechanism for catching new tweets

```
/** The actual Twitter stream. It's set up to collect raw JSON data */
private final TwitterStream twitterStream = new TwitterStreamFactory(
  new ConfigurationBuilder().setJSONStoreEnabled(true).build())
    .getInstance();
...
// The StatusListener is a twitter4j API that can be added to a stream,
// and will call a method every time a message is sent to the stream.
StatusListener listener = new StatusListener() {
  // The onStatus method is executed every time a new tweet comes in.
  public void onStatus(Status status) {
    ...
  }
}
...
// Set up the stream's listener (defined above), and set any necessary
// security information.
twitterStream.addListener(listener);
twitterStream.setOAuthConsumer(consumerKey, consumerSecret);
AccessToken token = new AccessToken(accessToken, accessTokenSecret);
twitterStream.setOAuthAccessToken(token);
```

JSON data

- JSON data is processed as an event and written to HDFS

```
public void onStatus(Status status) {
  // The EventBuilder is used to build an event using the headers and
  // the raw JSON of a tweet

  headers.put("timestamp", String.valueOf(
    status.getCreatedAt().getTime()));
  Event event = EventBuilder.withBody(
    DataObjectFactory.getRawJSON(status).getBytes(), headers);

  channel.processEvent(event);
}
```

What is Hive?

- HiveQL
  - SQL like interface
- Hive interpreter converts HiveQL to MapReduce code
- Returns results to the client

Hive details

- Schema on read
- Scalar types (int, float, double, boolean, string)
- Complex types (struct, map, array)
- Metastore contains table definitions
  - Allows queries to be data agnostic
  - Stored in a relational database
  - Similar to catalog tables in other DBs

Hive Serializers and Deserializers (SerDe)

- Instructs Hive on how to interpret data
- JSONSerDe

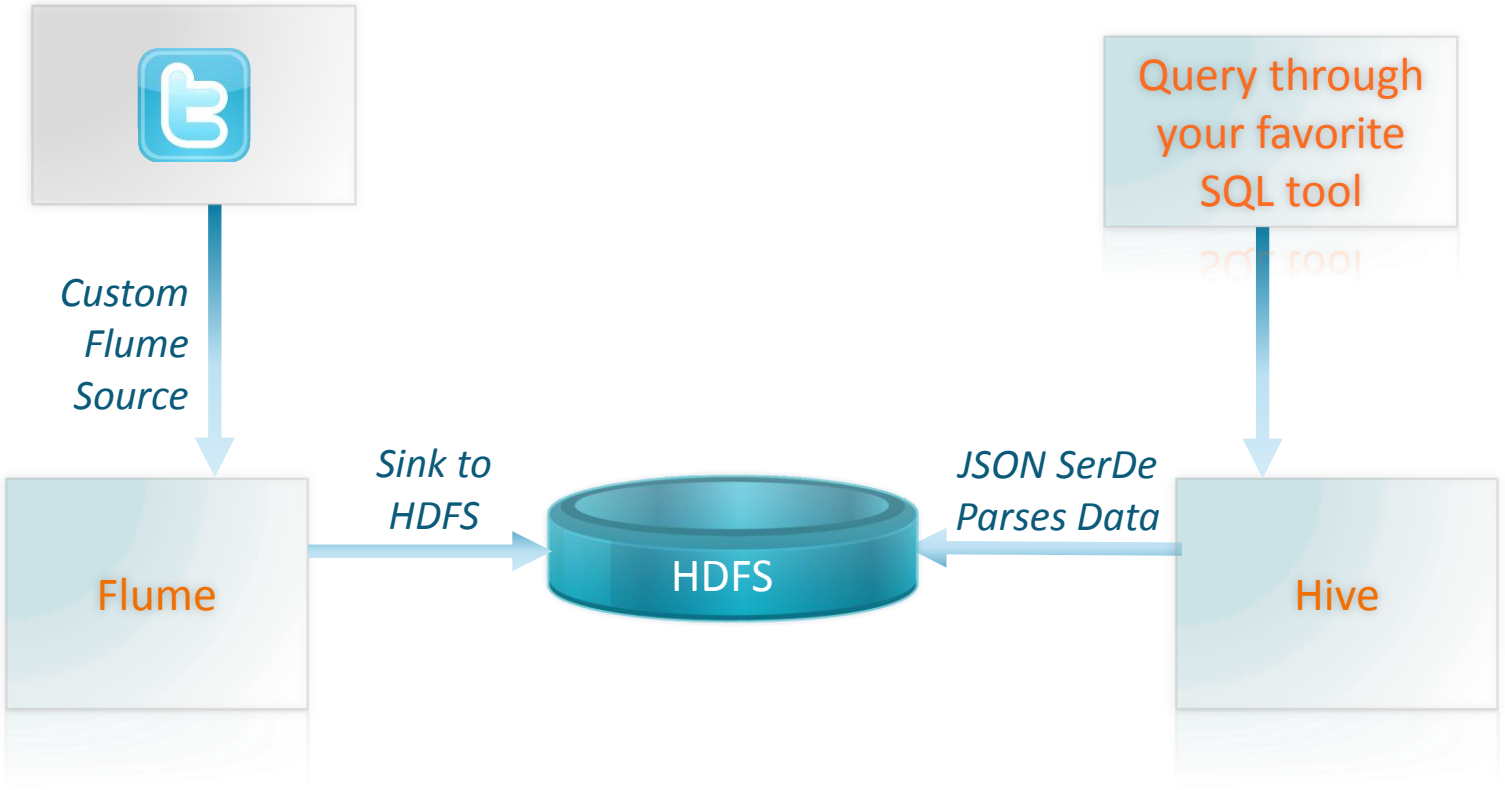
Hive Strenghts:

- Flexible in the data model
- Extendable format support

Analyzing Twitter data with Hadoop

PUTTING IT ALL TOGETHER

Architecture



Now We Can Start Asking Bigger Questions…

```
SELECT
  t.retweeted_screen_name,
  sum(retweets) AS total_retweets,
  count(*) AS tweet_count
FROM (SELECT
        retweeted_status.user.screen_name AS retweet_screen_name,
        retweeted_status.text,
        max(retweet_count) AS retweets
      FROM tweets
      GROUP BY
          retweeted_status.user.screen_name,
        retweeted_status.text) t
GROUP BY t.retweet_screen_name
ORDER BY total_retweets DESC
LIMIT 10;
```

Analyzing Twitter data with Hadoop

TEASER: FASTER HIVE? GO IMPALA!

Try it out yourself?

- Cloudera provides demo VMs
  - https://ccp.cloudera.com/display/SUPPORT/Cloudera+Manager+Free+Edition+Demo+VM
- More info and examples
  - http://blog.cloudera.com/

Beyond Big and Data

Prelude to a Philosophy of the BI Future

Lars Sjödin

Analyzing Twitter data with Hadoop

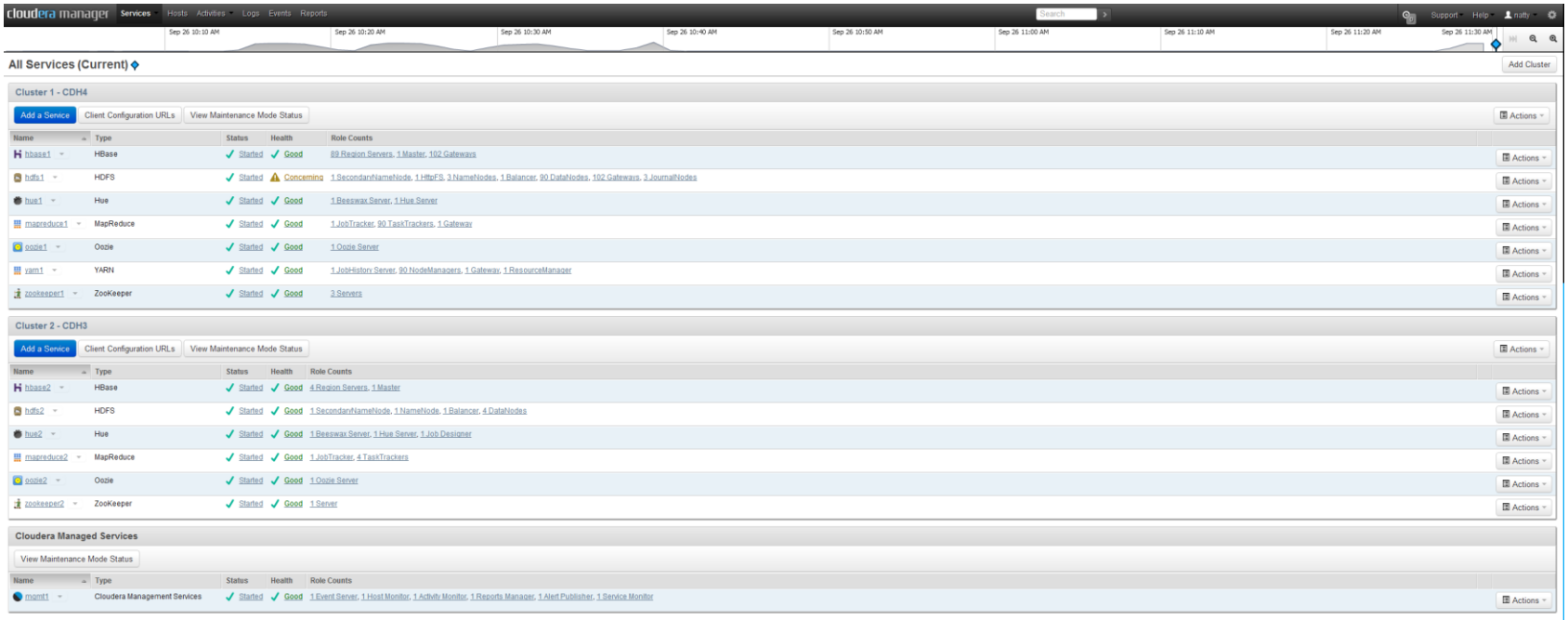
EXTRA SLIDES

NOTE: Hive is not a database

	RDBMS	Hive
Language	Generally >= SQL-92	Subset of SQL-92 plus Hive specific extensions
Update Capabilities	INSERT, UPDATE, DELETE	INSERT OVERWRITE no UPDATE, DELETE
Transactions	Yes	No
Latency	Sub-second	Minutes
Indexes	Yes	Yes
Data size	Terabytes	Petabytes

My personal preference to reduce complexity

- Cloudera Manager
  - https://ccp.cloudera.com/display/SUPPORT/Downloads
- Free up to 50 nodes



Analyzing Twitter data with Hadoop

JSON INTERLUDE

What is JSON?

- Complex, semi-structured data
- Based on JavaScript's data syntax
- Rich, nested data types:
  - number
  - string
  - Array
  - object
  - true, false
  - null

What is JSON?


```
{
  "retweeted_status": {
    "contributors": null,
    "text": "#Crowdsourcing – drivers already generate traffic data for your smartphone to suggest alternative routes when a road is clogged. #bigdata",
    "retweeted": false,
    "entities": {
      "hashtags": [
        {
          "text": "Crowdsourcing",
          "indices": [0, 14]
        },
        {
          "text": "bigdata",
          "indices": [129,137]
        }
      ],
      "user_mentions": []
    }
  }
}
```


Analyzing Twitter data with Hadoop

OOZIE: AUTOMATION

Oozie: everything in its right place

Oozie for partition management

- Once an hour, add a partition
- Takes advantage of advanced Hive functionality