THE BEST OF THE REST

Other relevant technologies, in brief

Impala

- Cloudera's alternative to Hive
- Massively parallel SQL engine on Hadoop
- Impala's always running, so you avoid the start-up costs when starting a Hive query
  - *Made for BI-style queries*
- Bottom line: Impala's often faster than Hive, but Hive offers more versatility
- Consider using Impala instead of Hive if you're using Cloudera

Accumulo

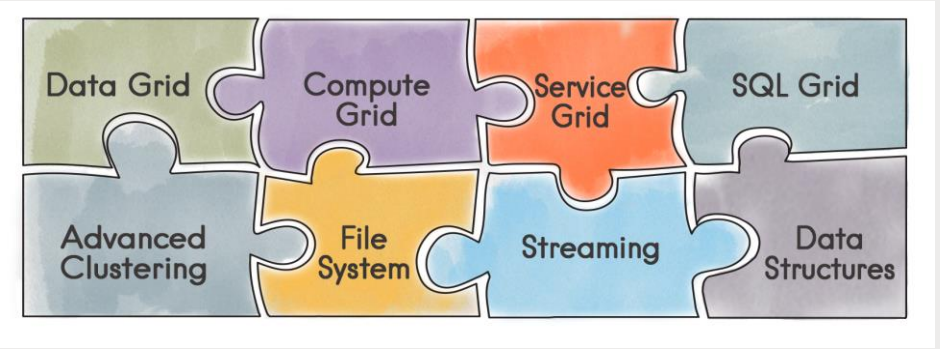
- Another BigTable clone (like HBase)
- But offers a better security model
  - *Cell-based access control*
- And server-side programming
- Consider it for your NoSQL needs if you have complex security requirements
  - *But make sure the systems that need to read this data can talk to it.*

Redis

- A distributed in-memory data store (like memcache)
- But it’s more than a cache!
- Good support for storing data structures
- Can persist data to disk
- Can be used as a data store and not just a cache
- Popular caching layer for web apps

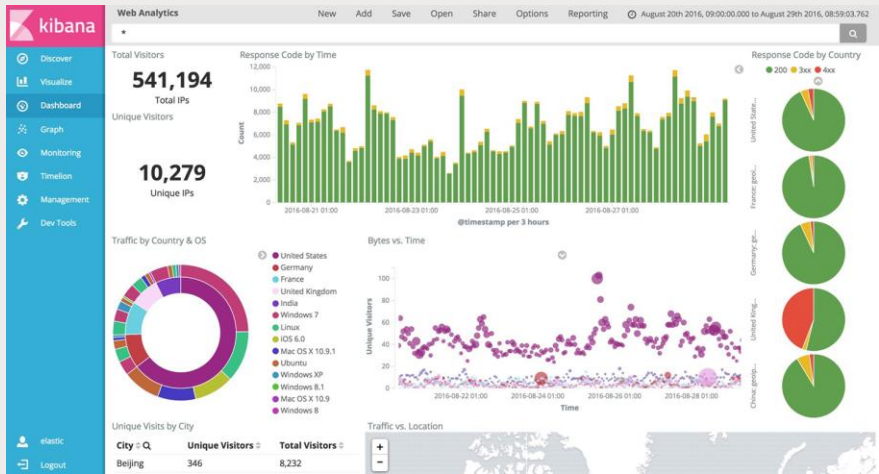
Ingite

- An “in-memory data fabric”
- Think of it as an alternative to Redis
- But it’s closer to a database
  - *ACID guarantees*
  - *SQL support*
  - *But it’s all done in-memory*



Elasticsearch

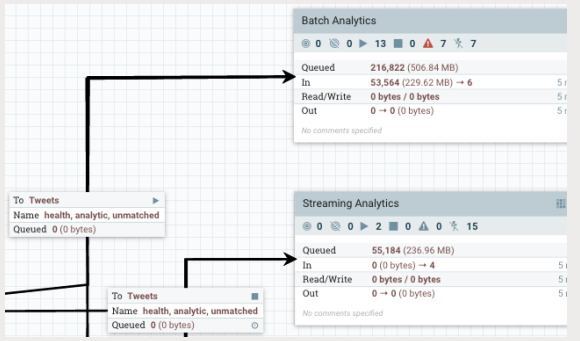
- A distributed document search and analytics engine
- Really popular
  - *Wikipedia, The Guardian, Stack Overflow, many more*
- Can handle things like real-time search-as-you-type
- When paired with Kibana, great for interactive exploration
- Amazon offers an Elasticsearch Service



Kinesis (and the AWS ecosystem)

- Amazon Kinesis is basically the AWS version of Kafka
- Amazon has a whole ecosystem of its own
  - *Elastic MapReduce (EMR)*
  - *S3*
  - *Elasticsearch Service / CloudSearch*
  - *DynamoDB*
  - *Amazon RDS*
  - *ElastiCache*
  - *AI / Machine Learning services*
- EMR in particular is an easy way to spin up a Hadoop cluster on demand

Apache NiFi



- Directed graphs of data routing
  - *Can connect to Kafka, HDFS, Hive*
- Web UI for designing complex systems
- Often seen in the context of IoT sensors, and managing their data
- Relevant in that it can be a streaming data source you'll see

Falcon

- A “data governance engine” that sits on top of Oozie
- Included in Hortonworks
- Like NiFi, it allows construction of data processing graphs
- But it’s really meant to organize the flow of your data within Hadoop

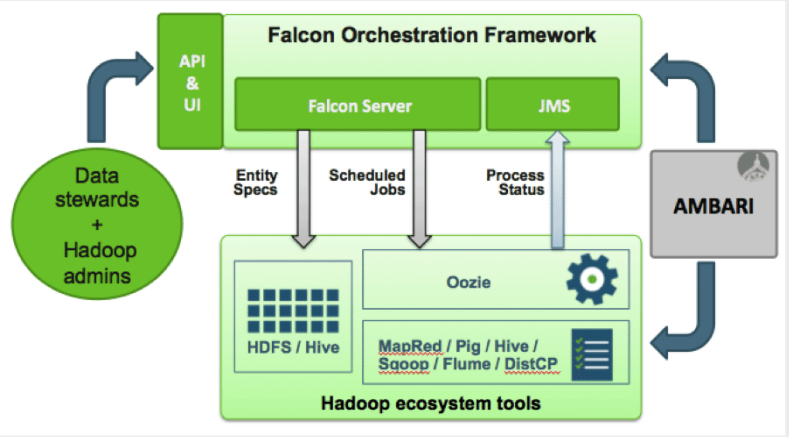
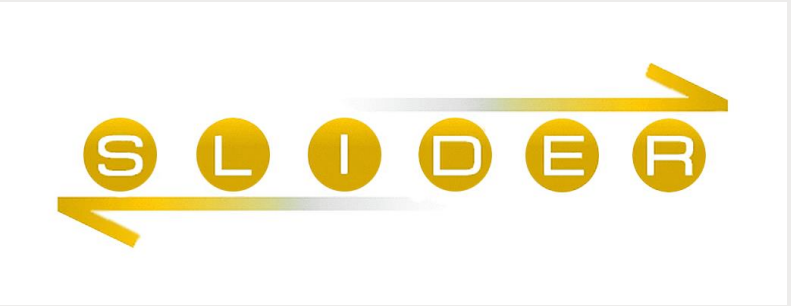


Image: falcon.apache.org

Apache Slider



- Deployment tool for apps on a YARN cluster
- Allows monitoring of your apps
- Allows growing or shrinking your deployment as it’s running
- Manages mixed configurations
- Start / stop applications on your cluster
- Incubating

And many more…

- Is your head spinning yet?