NIST system admin

Orchestrating

1a. system HW:

YARN 5 racked servers running Linux Ubuntu or CentOS on physical servers or Xen VMs. 3 copies striped over the chosen hask/key.

1b. system SW:

Puppet.

Base Hadoop distro selection on HW and OS. <http://wiki.apache.org/hadoop/Distributions%20and%20Commercial%20Support> Hyperclouds have resident Hadoop distros. These might be significantly faster to stand up and get running. Hortonworks has local office; check rolodex for local guy…

[tim] Using SSH for now. Remember to install Java 7 or 8. Be prepared to clean install the OS after you are ready to remove Hadoop.

1c. application SW.

2. plan for security, network, connectivity, v12n and management **--**

3. storage: Extract datasets into HDFS Transforming everything to XML.

Data.gov <http://www.data.gov>

Datahub <http://datahub.io/>

Facebook and <https://developers.facebook.com/>

Google trends and <https://developers.google.com/analytics/>

RSS or Atom feed from NYT, WaPost, or BBC

Twitter <https://twitter.com/?lang=en> and <https://dev.twitter.com/overview/documentation>

Looking for both batch and streaming APIs.

[Russell: document]

1. the data format and API for the ingested data,
2. all ETL transformation steps,
3. the data format stored in HDFS including hash/key.
4. decide now where the data sits prior to ingestion, including whether batch or pseudo live streaming.

Optionally,

[Russell: document]

1. how the data format in HDFS facilitates data discovery and pattern discovery.
2. list the SQL, noSQL, or MapReduce Java that you will run on the HDFS dataset. Hint: decide now where the data sits prior to ingestion, including whether batch or pseudo live streaming.

Hadoop runs standalone, single server, or fully distributed across servers based on the installation script.

Plan to use ssh login to every server for now; later, use REST or cloud broker to scale up to 1,000's of servers.

Plan as though you are downloading the SW and dataset(s) from github across the Internet.

[tim] management perspective for all six NIST datasets is the traditional experimental method. First, write some hypothesis in terms of metrics to be measured. Second, write the measurements taken during the actual measurement process (e.g., Boolean logic, probabilities, links). Third, write the anticipated NULL hypothesis model estimating the probability of detection plus probability of false alarm for all results. My point is that some isolated answer is not actionable until that answer is vetted by several causal or counterfactual explanations.

[tim] engineering perspective for all six NIST datasets follows.

1. First, focus on orchestrating system HW, system SW, and application SW.
2. Plan for security, network, connectivity, v12n,
3. Second, focus on ingesting & ETL'ing into HDFS datasets.
4. Third, actually perform experiments that end up with all six NIST datasets in a single HDFS.

Ignore data analysis until after stable and repeatable HDFS systems are available. Plan to run many experiments; push the envelope; fail early; measure performance; identify and document problems and bugs. Check bug lists.

The motivation for having orchestrators is best engineering practices. Some people will have expertise in a few areas, but often no one has expertise in all necessary areas. HW orchestration is one area (e.g., physical servers, admin VLAN, data VLAN, VM hypervisors, identity management incl federation, logging, auditing). System SW orchestration is another area (e.g., Linux distros, OpenStack, Java, containers; source or binaries on Internet or local repositories). Application orchestration is another area (e.g., download and install OS, containers, applications).

Data scientists usually assume someone else is responsible for the HW and SW. Those someones usually will not permit data scientists to change the system level HW or SW. Data scientists share; HW and SW admins keep everyone in their lanes.

Think of scripts as football cards to be traded.