Summary of 51 Uses and use in studying reference architecture

Government Operation (1-4): National Archives and Records Administration, Census Bureau

Maybe best captured by Bob Marcus's 10 different data scenarios. Adding recommender engines (#4) maybe easier for Netflix use case (#7)

Commercial (5-12): Finance in Cloud, Cloud Backup, Mendeley (Citations), Netflix, Web Search, Digital Materials, Cargo shipping (as in UPS)

Some of these are also perhaps best done from Bob Marcus's 10 use cases.

Netflix (#7) and Search (#8) can be done using open data sets

Cargo shipping (#10) is Internet of Things example; can be done with other streaming inputs

Defense (13-15): Sensors, Image surveillance, Situation Assessment

Important GIS/Image processing example. Could (all 3) be done as a disaster management (earthquake) scenario and linked to QuakeSim (#44)

Healthcare and Life Sciences (16-25): Medical records, Graph and Probabilistic analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity models, Biodiversity

Medical records (#16, 21, 22) could be hard due to usual privacy issues. Need discussion with Security&Privacy subgroup and submitters. Some issues are probed by Bob Marcus use cases

Pathology (#17) well developed by submitter with Hadoop. Good example

Imaging an important case (#18). Lots of images available. Should be able to design this nicely

Genomics (#19, 20) has plenty of available data and several developed projects to build on

Epidemiology (#23, 24) has well developed software at Virginia Tech. Work with submitters

Lifewatch (#25). Need to ask Yuri. Not clear if developed enough

Deep Learning and Social Media (26-31): Driving Car, Geolocate images/cameras, Twitter, Crowd Sourcing, Network Science, NIST benchmark datasets

Crowd sourcing (#29) vague; doubt if can be done well

NIST data (#31) could drive other applications. Could motivate a comparison project such as compare Hbase and Riak

Others (#26, 27, 28, 30) could be candidates. #26 requires GPU's and high performance clusters

The Ecosystem for Research (32-35): Metadata, Collaboration, Language Translation, Light source experiments

iRODS (#32) using "Apache Big Data" a good possible project as challenges clear as iRODS well developed and understood

#33 and #34 seem hard

Light sources (#35) comes back to #18

Astronomy and Physics (36-40): Sky Surveys compared to simulation, Large Hadron Collider at CERN, Belle Accelerator II in Japan

Astronomy (#36-38) is a good area and open data exists. There have been papers on astronomy in clouds. Non trivial but good project certainly possible

Particle Physics (#39, 40) should be possible using "fake data"

Earth, Environmental and Polar Science (41-50): Radar Scattering in Atmosphere, Earthquake, Ocean, Earth Observation, Ice sheet Radar scattering, Earth radar mapping, Climate simulation datasets, Atmospheric turbulence identification, Subsurface Biogeochemistry (microbes to watersheds), AmeriFlux and FLUXNET gas sensors

EISCAT (#41) and ENVRI (#42) require discussion with submitters. ENVRI seems attractive as a generic architecture that could be a good test for NIST reference architecture. #41 may be similar to #43

CReSIS (#43) can be set up with real data

QuakeSim (#44) related to imagery (#18, 35) and defense (#13-15). Project can be set up

NASA use of iRODS (#45) can be paired with #32.

MERRA (#46) needs to be discussed with NASA. Involves simulations (see #48)

Atmospheric turbulence related to Morris Riedel use case and image analysis (#17 #18) as looking for pathologies in turbulence (cf #17)

#48 exemplifies data from simulation seen in earlier examples such as #12 and #37. Good to have an example of this class

#49 seems not well developed yet. #50 is a good Internet of things case

Energy (51): Smart grid

Should have an Internet of Things example (see #10 #49). Can be done with other "Things" (sensors)

#52: Morris Riedel Use Case in looking for outliers

Indiana has NASA funded example here in hand. Should not be difficult to set up Morris's

Ogres as Summary of 51+1 use cases by characteristics

This follows Berkeley dwarfs (http://www.eecs.berkeley.edu/Pubs/TechRpts/2006/EECS-2006-183.html), NAS parallel benchmarks 9 https://www.nas.nasa.gov/publications/npb.html) and Linear Algebra Templates (http://www.netlib.org/linalg/html_templates/Templates.html). Work by Fox, Jha, Qiu

he purpose of Big Data Ogres is to discern commonalities and patterns across a broad range of seemingly different Big data applications, propose an initial structure to classify them, and help cluster some commonly found applications using structure. Note the Big Data Ogres, like the Berkeley Dwarfs are not orthogonal, nor exclusive, and thus do not constitute a formal taxonomy. Also we capture the

richness of Big data by including not just different parallel structures (as in 5th ogre below) but also important overall patterns. Big data is in its infancy without clear consensus as to important issues and so we propose an inclusive set of Ogres expecting that further discussion will refine them.

The first Ogre captures {\bf different analytical approaches} challenges. Some representative application classes are (i) Pleasingly Parallel -- as in Blast (over sequences), Protein docking (over proteins and docking sites), imagery (ii) Local Machine Learning -- ML or filtering pleasingly parallel as in bioimagery, radar (This contrasts with Global Machine Learning seen in LDA, Clustering etc. with parallel ML over nodes of system) (iii) Fusion: Knowledge discovery often involves fusion of multiple methods (ensemble methods one approach)

The second Ogre captures applications with {\bf important Data sources with distinctive features}, representative examples of the data sources include, (i) SQL based, (ii) NOSQL based, (iii) Set of Files (as managed in iRODS), (iv) Internet of Things, (v) Streaming and (vi) HPC simulations.

The third Ogre contains {\bf Distinctive System features}, and includes (i) Agents, as in epidemiology (swarm approaches) and (ii) GIS (Geographical Information Systems).

The forth Ogre builds upon the {\bf Problem Structure} of Big Data applications. For example, (i) Typical N points in a space; important differences between metric and non-metric spaces (ii) Maximum Likelihood, (iii) Chi-squared distributions, and (iv) Expectation Maximization (often method of Steepest descent).

The fifth Ogre organizes {\bf structure of the core analytics kernel} with representative examples (i) Recommender Systems (Collaborative Filtering) (ii) SVM and Linear Classifiers (Bayes, Random Forests), (iii) Outlier Detection (iORCA) (iv) Clustering (many methods), (v) PageRank, (vi) LDA (Latent Dirichlet Allocation), (vii) PLSI (Probabilistic Latent Semantic Indexing), (viii) SVD (Singular Value Decomposition), (ix) MDS (Multidimensional Scaling), (x) Graph Algorithms (seen in neural nets, search of RDF Triple stores), (xi) Neural Networks (Deep Learning), and (xii) Global Optimization (Variational Bayes)

A sixth Ogre could describe the Bob Marcus examples below.

Bob Marcus Use Cases

Ten possible simple core use cases with examples:

1. Multiple users performing interactive queries and updates on a database with basic availability and eventual consistency (BASE)

Big Data File Systems as a data resource for batch and interactive queries

NoSQL (and NewSQL) DBs as operational databases for large-scale updates and queries

NoSQL DBs for storing diverse data types NoSQL DBs for storing diverse data types

Databases optimized for rapid updates and retrieval (e.g. in memory or SSD)

2. Perform real time analytics on data source streams and notify users when specified events occur

Operations Analysis

Stream Processing and ETL

Real Time Analytics (e.g. Complex Event Processing)

3. Move data from external data sources into a highly horizontally scalable data store, transform it using highly horizontally scalable processing (e.g. Map-Reduce), and return it to the horizontally scalable data store (ELT)

Data input and output to Big Data File System (ETL, ELT)

Stream Processing and ETL

4. Perform batch analytics on the data in a highly horizontally scalable data store using highly horizontally scalable processing (e.g Map-Reduce) with a user-friendly interface (e.g. SQL like)

Big Data Exploration

Data Warehouse Augmentation

Big Data File Systems as a data resource for batch and interactive queries

5. Perform interactive analytics on data in analytics-optimized database

Big Data Exploration

Data Warehouse Augmentation

Databases optimized for complex ad hoc queries

Databases optimized for rapid updates and retrieval (e.g. in memory or SSD)

Big Data File Systems used as a data resource for interactive queries

6. Visualize data extracted from horizontally scalable Big Data score

Big Data Exploration

Visualization Tools for End-Users

7. Move data from a highly horizontally scalable data store into a traditional Enterprise Data Warehouse

Data input and output to Big Data File System (ETL, ELT)

Data exported to Databases from Big Data File System

Data Warehouse Augmentation

8. Extract, process, and move data from data stores to archives

9. Combine data from Cloud databases and on premise data stores for analytics, data mining, and/or machine learning

10. Orchestrate multiple sequential and parallel data transformations and/or analytic processing using a workflow manager

Big Data Exploration

Enhanced 360º View of the Customer

Security/Intelligence Extension