Big Data Use Cases and Requirements

Co-Chairs:

Geoffrey Fox, Indiana University (gcfexchange@gmail.com)

Ilkay Altintas, UCSD/SDSC (altintas@sdsc.edu)
Requirements and Use Case Subgroup

The focus is to form a community of interest from industry, academia, and government, with the goal of developing a consensus list of Big Data requirements across all stakeholders. This includes gathering and understanding various use cases from diversified application domains.

Tasks
• Gather use case input from all stakeholders
• Derive Big Data requirements from each use case.
• Analyze/prioritize a list of challenging general requirements that may delay or prevent adoption of Big Data deployment
• Work with Reference Architecture to validate requirements and reference architecture
• Develop a set of general patterns capturing the “essence” of use cases (to do)
<table>
<thead>
<tr>
<th>Use Case Title</th>
<th>Vertical (area)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Author/Company/Email</td>
<td>Actors/Stakeholders and their roles and responsibilities</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Goals</th>
</tr>
</thead>
<tbody>
<tr>
<td>Use Case Description</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Current Solutions</th>
<th>Compute(System)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Storage</td>
</tr>
<tr>
<td></td>
<td>Networking</td>
</tr>
<tr>
<td></td>
<td>Software</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data Characteristics</th>
<th>Data Source (distributed/centralized)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Volume (size)</td>
</tr>
<tr>
<td></td>
<td>Velocity (e.g. real time)</td>
</tr>
<tr>
<td></td>
<td>Variety (multiple datasets, mashup)</td>
</tr>
<tr>
<td></td>
<td>Variability (rate of change)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data Science (collection, curation, analysis, action)</th>
<th>Veracity (Robustness Issues, semantics)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Visualization</td>
</tr>
<tr>
<td></td>
<td>Data Quality (syntax)</td>
</tr>
<tr>
<td></td>
<td>Data Types</td>
</tr>
<tr>
<td></td>
<td>Data Analytics</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Big Data Specific Challenges (Gaps)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Big Data Specific Challenges in Mobility</td>
</tr>
<tr>
<td>Security &amp; Privacy Requirements</td>
</tr>
<tr>
<td>Highlight issues for generalizing this use case (e.g. for ref. architecture)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>More Information (URLs)</th>
</tr>
</thead>
</table>

Note: No proprietary or confidential information should be included
ADD picture of operation or data architecture of application below table

OX - March, 2014
51 Detailed Use Cases: Many TB’s to Many PB’s

- **Government Operation:** National Archives and Records Administration, Census Bureau
- **Commercial:** Finance in Cloud, Cloud Backup, Mendeley (Citations), Netflix, Web Search, Digital Materials, Cargo shipping (as in UPS)
- **Defense:** Sensors, Image surveillance, Situation Assessment
- **Healthcare and Life Sciences:** Medical records, Graph and Probabilistic analysis, Pathology, Bioimaging, Genomics, Epidemiology, People Activity models, Biodiversity
- **Deep Learning and Social Media:** Driving Car, Geolocate images/cameras, Twitter, Crowd Sourcing, Network Science, NIST benchmark datasets
- **The Ecosystem for Research:** Metadata, Collaboration, Language Translation, Light source experiments
- **Astronomy and Physics:** Sky Surveys compared to simulation, Large Hadron Collider at CERN, Belle Accelerator II in Japan
- **Earth, Environmental and Polar Science:** 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
- **Energy:** Smart grid

Next step involves matching extracted requirements and reference architecture. Alternatively develop a set of general patterns capturing the “essence” of use cases.
Some Trends

• Practitioners consider themselves Data Scientists
• Images are a major source of Big Data
  – Radar
  – Light Synchrotrons
  – Phones
  – Bioimaging

• Hadoop and HDFS dominant
• Business – main emphasis at NIST – interested in analytics and assume HDFS
• Academia also extremely interested in data management
• Clouds v. Grids

Source: KPCB estimates based on publicly disclosed company data.
Example Use Case I: Summary of Genomics

• **Application:** 19: NIST Genome in a Bottle Consortium
  – integrates data from multiple sequencing technologies and methods
  – develops highly confident characterization of whole human genomes as reference materials,
  – develops methods to use these Reference Materials to assess performance of any genome sequencing run.

• **Current Approach:**
  – The storage of ~40TB NFS at NIST is full; there are also PBs of genomics data at NIH/NCBI.
  – Use Open-source sequencing bioinformatics software from academic groups on a 72 core cluster at NIST supplemented by larger systems at collaborators.

• **Futures:**
  – DNA sequencers can generate ~300GB compressed data/day which volume has increased much faster than Moore’s Law.
  – Future data could include other ‘omics’ measurements, which will be even larger than DNA sequencing. Clouds have been explored.
Example Use Case II: Census Bureau Statistical Survey Response Improvement (Adaptive Design)

- **Application:** Survey costs are increasing as survey response declines.
  - Uses advanced “recommendation system techniques” that are open and scientifically objective
  - Data mashed up from several sources and historical survey para-data (administrative data about the survey) to drive operational processes
  - The end goal is to increase quality and reduce the cost of field surveys

- **Current Approach:**
  - ~1PB of data coming from surveys and other government administrative sources.
  - Data can be streamed with approximately 150 million records transmitted as field data streamed continuously, during the decennial census.
  - All data must be both confidential and secure.
  - All processes must be auditable for security and confidentiality as required by various legal statutes.
  - Data quality should be high and statistically checked for accuracy and reliability throughout the collection process.

- **Futures:**
  - Analytics needs to be developed which give statistical estimations that provide more detail, on a more near real time basis for less cost.
  - The reliability of estimated statistics from such “mashed up” sources still must be evaluated.
Deep Learning and Social Media

Example Use Case III: 26: Large-scale Deep Learning

• Application: 26: Large-scale Deep Learning
  – Large models (e.g., neural networks with more neurons and connections) combined with large datasets are increasingly the top performers in benchmark tasks for vision, speech, and Natural Language Processing.
  – One needs to train a deep neural network from a large (>>1TB) corpus of data (typically imagery, video, audio, or text).
  – Such training procedures often require customization of the neural network architecture, learning criteria, and dataset pre-processing.
  – In addition to the computational expense demanded by the learning algorithms, the need for rapid prototyping and ease of development is extremely high.

• Current Approach:
  – The largest applications so far are to image recognition and scientific studies of unsupervised learning with 10 million images and up to 11 billion parameters on a 64 GPU HPC Infiniband cluster.
  – Both supervised (using existing classified images) and unsupervised applications investigated.

• Futures:
  – Large datasets of 100TB or more may be necessary in order to exploit the representational power of the larger models.
  – Training a self-driving car could take 100 million images at megapixel resolution.
  – Deep Learning shares many characteristics with the broader field of machine learning. The paramount requirements are high computational throughput for mostly dense linear algebra operations, and extremely high productivity for researcher exploration.
  – One needs integration of high performance libraries with high level (python) prototyping environments.
Example Use Case IV: 
EISCAT 3D incoherent scatter radar system

• Application: *EISCAT 3D incoherent scatter radar system*
  – EISCAT: European Incoherent Scatter Scientific Association
  – Research on the lower, middle and upper atmosphere and ionosphere using the incoherent scatter radar technique.
  – This technique is the most powerful ground-based tool for these research applications.
  – EISCAT studies instabilities in the ionosphere, as well as investigating the structure and dynamics of the middle atmosphere. It is also a diagnostic instrument in ionospheric modification experiments with addition of a separate Heating facility.
  – Currently EISCAT operates 3 of the 10 major incoherent radar scattering instruments worldwide with its facilities in in the Scandinavian sector, north of the Arctic Circle.

• Current Approach:
  – The current old EISCAT radar generates terabytes per year rates and no present special challenges.

• Futures:
  – The next generation radar, EISCAT_3D, will consist of a core site with a transmitting and receiving radar arrays and four sites with receiving antenna arrays at some 100 km from the core.
  – The fully operational 5-site system will generate several thousand times data of current EISCAT system with 40 PB/year in 2022 and is expected to operate for 30 years.
  – EISCAT 3D data e-Infrastructure plans to use the high performance computers for central site data processing and high throughput computers for mirror sites data processing.
  – Downloading the full data is not time critical, but operations require real-time information about certain pre-defined events to be sent from the sites to the operation center and a real-time link from the operation center to the sites to set the mode of radar operation on with immediate action.
Energy

Example Use Case V: Consumption forecasting in Smart Grids

• **Application:** *51: Consumption forecasting in Smart Grids*
  - Predict energy consumption for customers, transformers, sub-stations and the electrical grid service area using smart meters providing measurements every 15-mins at the granularity of individual consumers within the service area of smart power utilities.
  - Combine Head-end of smart meters (distributed), Utility databases (Customer Information, Network topology; centralized), US Census data (distributed), NOAA weather data (distributed), Micro-grid building information system (centralized), Micro-grid sensor network (distributed).
  - This generalizes to real-time data-driven analytics for time series from cyber physical systems

• **Current Approach:**
  - GIS based visualization.
  - Data is around 4 TB a year for a city with 1.4M sensors in Los Angeles.
  - Uses R/Matlab, Weka, Hadoop software.
  - Significant privacy issues requiring anonymization by aggregation.
  - Combine real time and historic data with machine learning for predicting consumption.

• **Futures:**
  - Wide spread deployment of Smart Grids with new analytics integrating diverse data and supporting curtailment requests. Mobile applications for client interactions.
**Example Use Case VI: Pathology Imaging**

- **Application:** *Pathology Imaging/ Digital Pathology II*
- **Current Approach:**
  - 1GB raw image data + 1.5GB analytical results per 2D image.
  - MPI for image analysis; MapReduce + Hive with spatial extension on supercomputers and clouds.
  - GPU's used effectively. Figure below shows the architecture of Hadoop-GIS, a spatial data warehousing system over MapReduce to support spatial analytics for analytical pathology imaging.

- **Futures:** Recently, 3D pathology imaging is made possible through 3D laser technologies or serially sectioning hundreds of tissue sections onto slides and scanning them into digital images. Segmenting 3D microanatomic objects from registered serial images could produce tens of millions of 3D objects from a single image. This provides a deep “map” of human tissues for next generation diagnosis. 1TB raw image data + 1TB analytical results per 3D image and 1PB data per moderated hospital per year.

[Architecture of Hadoop-GIS, a spatial data warehousing system over MapReduce to support spatial analytics for analytical pathology imaging]
Example Use Case VII: Metagenomics

- **Application:** *20: Comparative analysis for metagenomes and genomes*
  - Given a metagenomic sample, (1) determine the community composition in terms of other reference isolate genomes, (2) characterize the function of its genes, (3) begin to infer possible functional pathways, (4) characterize similarity or dissimilarity with other metagenomic samples, (5) begin to characterize changes in community composition and function due to changes in environmental pressures, (6) isolate sub-sections of data based on quality measures and community composition.

- **Current Approach:**
  - Integrated comparative analysis system for metagenomes and genomes, front ended by an interactive Web UI with core data, backend precomputations, batch job computation submission from the UI.
  - Provide interface to standard bioinformatics tools (BLAST, HMMER, multiple alignment and phylogenetic tools, gene callers, sequence feature predictors...).

- **Futures:**
  - Management of heterogeneity of biological data is currently performed by RDMS (Oracle). Unfortunately, it does not scale for even the current volume 50TB of data.
  - NoSQL solutions aim at providing an alternative but unfortunately they do not always lend themselves to real-time interactive use, rapid and parallel bulk loading, and sometimes have issues regarding robustness.
Example Use Case VIII: Consumer photography

• Application: 27: Organizing large-scale, unstructured collections of consumer photos
  – Produce 3D reconstructions of scenes using collections of millions to billions of consumer images, where neither the scene structure nor the camera positions are known a priori.
  – Use resulting 3d models to allow efficient browsing of large-scale photo collections by geographic position.
  – Geolocate new images by matching to 3d models. Perform object recognition on each image. 3d reconstruction posed as a robust non-linear least squares optimization problem where observed relations between images are constraints and unknowns are 6-d camera pose of each image and 3-d position of each point in the scene.

• Current Approach:
  – Hadoop cluster with 480 cores processing data of initial applications.
  – Note over 500 billion images on Facebook and over 5 billion on Flickr with over 500 million images added to social media sites each day.
**Futures:**

- Need many analytics including feature extraction, feature matching, and large-scale probabilistic inference, which appear in many or most computer vision and image processing problems, including recognition, stereo resolution, and image denoising.
- Need to visualize large-scale 3-d reconstructions, and navigate large-scale collections of images that have been aligned to maps.
Example Use Case IX: Twitter Data

- **Application:** *28: Truthy: Information diffusion research from Twitter Data*
  - Understanding how communication spreads on socio-technical networks.
  - Detecting potentially harmful information spread at the early stage (e.g., deceiving messages, orchestrated campaigns, untrustworthy information, etc.)

- **Current Approach:**
  - 1) Acquisition and storage of a large volume (30 TB a year compressed) of continuous streaming data from Twitter (~100 million messages/day, ~500GB data/day increasing);
  - 2) near real-time analysis of such data, for anomaly detection, stream clustering, signal classification and online-learning;
  - 3) data retrieval, big data visualization, data-interactive Web interfaces, public API for data querying. Use Python/SciPy/NumPy/MPI for data analysis. Information diffusion, clustering, and dynamic network visualization capabilities already exist

- **Futures:**
  - Truthy plans to expand incorporating Google+ and Facebook.
  - Need to move towards Hadoop/IndexedHBase & HDFS distributed storage.
  - Use Redis as an in-memory database to be a buffer for real-time analysis.
  - Need streaming clustering, anomaly detection and online learning.
<table>
<thead>
<tr>
<th>No.</th>
<th>Use Case</th>
<th>Volume</th>
<th>Velocity</th>
<th>Variety</th>
<th>Software</th>
<th>Analytics</th>
</tr>
</thead>
<tbody>
<tr>
<td>23</td>
<td>M0172 World Population Scale Epidemiological Study</td>
<td>100TB</td>
<td>Data feeding into the simulation is small but real time data generated by simulation is massive.</td>
<td>Can be rich with various population activities, geographical, socio-economic, cultural variations</td>
<td>Charm++, MPI</td>
<td>Simulations on a Synthetic population</td>
</tr>
<tr>
<td>24</td>
<td>M0173 Social Contagion Modeling for Planning</td>
<td>10s of TB per year</td>
<td>During social unrest events, human interactions and mobility leads to rapid changes in data; e.g., who follows whom in Twitter.</td>
<td>Data fusion a big issue. How to combine data from different sources and how to deal with missing or incomplete data?</td>
<td>Specialized simulators, open source software, and proprietary modeling environments. Databases.</td>
<td>Models of behavior of humans and hard infrastructures, and their interactions. Visualization of results</td>
</tr>
<tr>
<td>25</td>
<td>M0141 Biodiversity and LifeWatch</td>
<td>N/A</td>
<td>Real time processing and analysis in case of the natural or industrial disaster</td>
<td>Rich variety and number of involved databases and observation data</td>
<td>RDMS</td>
<td>Requires advanced and rich visualization</td>
</tr>
<tr>
<td>26</td>
<td>M0136 Large-scale Deep Learning</td>
<td>Current datasets typically 1 to 10 TB. Training a self-driving car could take 100 million images.</td>
<td>Much faster than real-time processing is required. For autonomous driving need to process 1000's high-resolution (6 megapixels or more) images per second.</td>
<td>Neural Net very heterogeneous as it learns many different features</td>
<td>In-house GPU kernels and MPI-based communication developed by Stanford. C++/Python source.</td>
<td>Small degree of batch statistical pre-processing; all other data analysis is performed by the learning algorithm itself.</td>
</tr>
<tr>
<td>27</td>
<td>M0171 Organizing large-scale image collections</td>
<td>500+ billion photos on Facebook, 5+ billion photos on Flickr.</td>
<td>Over 500M images uploaded to Facebook each day</td>
<td>Images and metadata including EXIF tags (focal distance, camera type, etc.), Hadoop Map-reduce, simple hand-written multithreaded tools (ssh and sockets for communication)</td>
<td>Hadoop, Redis for data management. Python: Robust non-linear least squares optimization problem. Support Vector Machine</td>
<td>Robust non-linear least squares optimization problem. Support Vector Machine</td>
</tr>
<tr>
<td>28</td>
<td>M0160 Truthy</td>
<td>30TB/year compressed data</td>
<td>Near real-time data storage, querying &amp; analysis</td>
<td>Schema provided by social media data source. Currently using Twitter only. Planned to expand</td>
<td>Hadoop IndexedHBase &amp; HDFS. Hadoop, Hive, Redis for data management. Python: Anomaly detection, stream clustering, signal classification and online-learning; Information diffusion,</td>
<td></td>
</tr>
</tbody>
</table>
Requirements Gathering

• Data sources
  – data size, file formats, rate of grow, at rest or in motion, etc.

• Data lifecycle management
  – curation, conversion, quality check, pre-analytic processing, etc.

• Data transformation
  – data fusion/mashup, analytics

• Capability infrastructure
  – software tools, platform tools, hardware resources such as storage and networking

• Security & Privacy; and data usage
  – processed results in text, table, visual, and other formats

A total of 437 specific requirements under 35 high-level generalized requirement summaries.
Interaction Between Subgroups

Due to time constraints, activities were carried out in parallel.
Reference Architecture

- Multiple stacks of technologies
  - Open and Proprietary
- Provide example stacks for different applications
- Come up with usage patterns and best practices
Next Steps

• Approach for RDA to implement use cases and NBD to identify abstract interface
• Planning for implementation of usecases
  – Resource availability
  – Application-specific support
  – Computation and storage leverage
• Multiple potential directions
  – Prioritization is one of the goals for this meeting.
Key Links

• Use cases listing:
  http://bigdatawg.nist.gov/usecases.php

• Latest version of the document (Dated Oct 12, 2013):
  http://bigdatawg.nist.gov/_uploadfiles/M0245_v5_6066621242.docx