**July 16 Agenda**

**Review charter and NIST guidelines including timeline**

**Brainstorm about work of Requirements WG**

What to do and how to do it?

**What To Do**

**Collect Use Cases**

Specific examples

Generalization to similar use cases

Need to define format of a use case so can generate abstractions

**Use cases linked to abstractions/generalizations of features as below**

Refine abstractions and link to use cases

Network, Storage, Compute needs

Infrastructure/Architecture Requirement (NoSQL v SQL, MapReduce v MPI, Clouds v HPC...)

Data size v Compute Size

Data centralized or distributed

See Bob Marcus document for more examples

Algorithms needed; Implementation needed (e.g. parallel or GPU versions)

Security&Privacy

Training or Expertise needed

**How to do it or Process to follow**

Gather documents and URL's/Citations

Agree on approach; formats & methodology of giving input

Specify in WG meeting

Edit document on Google docs

Upload to <http://bigdatawg.nist.gov/show_InputDoc2.php>

Interactions with other WG

Contact other individuals and organizations that can contribute use cases etc.

**Agree on what we will do and how we will do it.**

Participant Profile

Use Case Template

Uses Cases

**Agree on collaborative sharing of information**

**Note Initial Information for Requirements WG prepared as something to discuss by Fox**

**Working Group Profiles**

**Geoffrey Fox co-chair.** Professor in School of Informatics and Computing at Indiana University Bloomington within Informatics division. Current research (<http://www.infomall.org/> or <https://portal.futuregrid.org/> ) is in Cloud Computing and its applications (Bioinformatics, Sensor Clouds, Earthquake and Ice-sheet Science, and Particle Physics). I am working as part of data science program with possible online curriculum as at <http://x-informatics.appspot.com/preview>.

**Add yourself**

**-----------------------------------------------------------------------------------------**

**Use Case Discussion**

Upload detailed use cases to <http://bigdatawg.nist.gov/upload.php> specifying UseCases as subgroup

**Particle Physics: Analysis of LHC Data (Discovery of Higgs particle)**

**Goal:** Understanding of properties of fundamental particles

**Data Source and Nature:** CERN LHC Accelerator and Monte Carlo producing events describing particle-apparutus interaction. Processed information defines physics properties of events (lists of particles with type and momenta)

**Data Size:** 15 Petabytes per year

**Analysis/Compute Model:** 200,000 cores running “continuously” arranged in 3 tiers (CERN, “Continents/Countries”. “Universities”). Uses “High Throughput Computing” (Pleasing parallel). Can use commercial or academic clouds but analysis infrastructure built before clouds available

**Data Analytics:** Initial analysis is processing of experimental data specific to each experiment (ALICE, ATLAS, CMS, LHCb) producing summary information. Second step in analysis uses “exploration” (histograms, scatter-plots) with model fits. Substantial Monte-Carlo computations to estimate analysis quality.

**More Information:** <http://grids.ucs.indiana.edu/ptliupages/publications/Where%20does%20all%20the%20data%20come%20from%20v7.pdf>

**Netflix Movie Service**

**Goal:** Allow streaming of user selected movies to satisfy multiple objectives -- especially retaining subscribers.

**Data Source and Nature:** Digital movies stored in cloud with metadata; user profiles and rankings for small fraction of movies for each user.

**Data size:** Summer 2012. 25 million subscribers; 4 million ratings per day; 3 million searches per day; 1 billion hours streamed in June 2012

**Compute Model:** Amazon Web Services AWS. Uses Cassandra NoSQL technology

**Data Analytics:** Recommender systems and streaming video delivery. Recommender systems use logistic/linear regression, elastic nets, matrix factorization, clustering, latent Dirichlet allocation, association rules, gradient boosted decision trees and others. Winner of Netflix competition (to improve ratings by 10%) combined over 100 different algorithms.

**More Information:** <http://www.slideshare.net/xamat/building-largescale-realworld-recommender-systems-recsys2012-tutorial>

**Add your examples**

**-----------------------------------------------------------------------------------------**

**Use Case Template or Best Practices in Specifying use cases**

**see specific cases above**

**-----------------------------------------------------------------------------------------**