**NBD(NIST Big Data) Requirements WG Use Case Template**

|  |  |  |  |
| --- | --- | --- | --- |
| **Use Case Title** | | Web Search (Bing, Google, Yahoo..) | |
| **Vertical (area)** | | Commercial Cloud Consumer Services | |
| **Author/Company/Email** | | Geoffrey Fox, Indiana University gcf@indiana.edu | |
| **Actors/Stakeholders and their roles and responsibilities** | | Owners of web information being searched; search engine companies; advertisers; users | |
| **Goals** | | Return in ~0.1 seconds, the results of a search based on average of 3 words; important to maximize “precisuion@10”; number of great responses in top 10 ranked results | |
| **Use Case Description** | | .1) Crawl the web; 2) Pre-process data to get searchable things (words, positions); 3) Form Inverted Index mapping words to documents; 4) Rank relevance of documents: PageRank; 5) Lots of technology for advertising, “reverse engineering ranking” “preventing reverse engineering”; 6) Clustering of documents into topics (as in Google News) 7) Update results efficiently | |
| **Current**  **Solutions** | **Compute(System)** | | Large Clouds |
| **Storage** | | Inverted Index not huge; crawled documents are petabytes of text – rich media much more |
| **Networking** | | Need excellent external network links; most operations pleasingly parallel and I/O sensitive. High performance internal network not needed |
| **Software** | | MapReduce + Bigtable; Dryad + Cosmos. Final step essentially a recommender engine |
| **Big Data  Characteristics** | **Data Source (distributed/centralized)** | | Distributed web sites |
| **Volume (size)** | | **45B web pages total, 500M photos uploaded each day, 100 hours of video uploaded to YouTube each minute** |
| **Velocity**  **(e.g. real time)** | | **Data continually updated** |
| **Variety**  **(multiple datasets, mashup)** | | **Rich set of functions. After processing, data similar for each page (except for media types)** |
| **Variability (rate of change)** | | **Average page has life of a few months** |
| **Big Data Science (collection, curation,**  **analysis,**  **action)** | **Veracity (Robustness Issues)** | | **Exact results not essential but important to get main hubs and authorities for search query** |
| **Visualization** | | **Not important although page lay out critical** |
| **Data Quality** | | **A lot of duplication and spam** |
| **Data Types** | | **Mainly text but more interest in rapidly growing image and video** |
| **Data Analytics** | | **Crawling; searching including topic based search; ranking; recommending** |
| **Big Data Specific Challenges (Gaps)** | | Search of “deep web” (information behind query front ends)  Ranking of responses sensitive to intrinsic value (as in Pagerank) as well as advertising value  Link to user profiles and social network data | |
| **Big Data Specific Challenges in Mobility** | | Mobile search must have similar interfaces/results | |
| **Security & Privacy**  **Requirements** | | Need to be sensitive to crawling restrictions. Avoid Spam results | |
| **Highlight issues for generalizing this use case (e.g. for ref. architecture)** | | Relation to Information retrieval such as search of scholarly works. | |
| **More Information (URLs)** | | http://www.slideshare.net/kleinerperkins/kpcb-internet-trends-2013  http://webcourse.cs.technion.ac.il/236621/Winter2011-2012/en/ho\_Lectures.html  http://www.ifis.cs.tu-bs.de/teaching/ss-11/irws  http://www.slideshare.net/beechung/recommender-systems-tutorialpart1intro  http://www.worldwidewebsize.com/ | |
| **Note:** <additional comments> | | | |