Introduction to Big Data & Basic Data Analysis

Freddy Wetjen,
National Library of Norway.

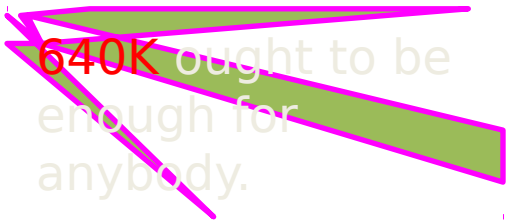
Big Data EveryWhere!

- Lots of data may be collected and warehoused
  - Web data, e-commerce
  - purchases at department/ grocery stores
  - Bank/Credit Card transactions
  - Social Network
  - Mobile usage & tracing

Nasjonalbiblioteket

How much data?

- Google processes 20 PB a day (2008)
- Wayback Machine has 3 PB + 100 TB/month (3/2009)
- Facebook has 2.5 PB of user data + 15 TB/day (4/2009)
- eBay has 6.5 PB of user data + 50 TB/day (5/2009)
- CERN’s Large Hydron Collider (LHC) generates 15 PB a year



Nasjonalbiblioteket

Type of Data

- Relational Data (Tables/Transaction/Legacy Data)
- Text Data (Web)
- Semi-structured Data (XML)
- “Record” structured (programming language)
- Graph Data
  - Social Network, Semantic Web (RDF), …
  - Usage data
- Streaming Data
  - You can only scan the data once

Nasjonalbiblioteket

What to do with these data?

- Aggregation and Statistics
  - Data warehouse and OLAP
- Indexing, Searching, and Querying
  - Keyword based search
  - Pattern matching (XML/RDF)
- Knowledge discovery
  - Data Mining
  - Statistical Modeling
  - Reasoning?

Nasjonalbiblioteket

Examples of Big data applications

Analysis of customer usage patterns (Netflix). Finding customer habits and do recommandations.

Data warehousing.

Seismic data analysis

Creating and connecting large collections of mixed media.

Gaining knowledge and experience across large heterogenous datasets.

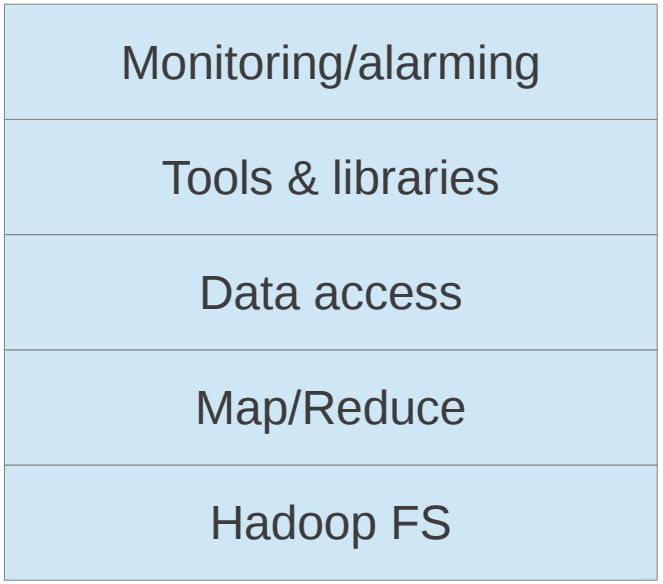
Knowledge based reasoning after information extraction.

"The problem is to be able to connect the large number of (different types of) dots to understand the pattern".

Nasjonalbiblioteket

Google Hadoop;Big data enabler

- GFS (Global File System)
- Language for expressing distributed processing
- Other tools/Frameworks



Hadoop world

Nasjonalbiblioteket

Global File System (GFS); Examplified with Hadoop

- Scalable open source framework for distributed storage of data.
- Scalable to petabytes and 1000's of nodes; heavily paralizable
- Started out from google...
- Buildt on top of commodity filesystems.
- Fault tolerant,scalable, in use...

Benefits of a Hadoop Global file system

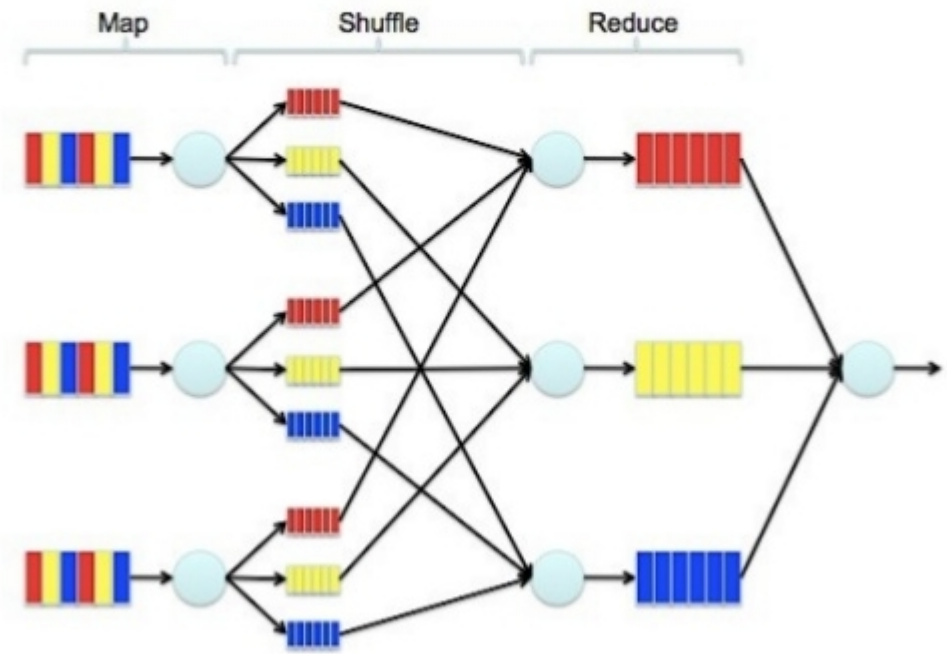
- All data available from a single mount point.
- Single point of management of filesystem.
- Single point of job management.
- Security,scalabililty and redundancy buildt into the system

Nasjonalbiblioteket

Map Reduce;Expressing distributed processing for Hadoop

- Two steps in expressing Processing; Map and Reduce.
- Map step to express the processing for each data unit.
- Reduce step to “coordinate” the different result sets.
- Specialized language for expressing this MapReduce → Pig
- Java based framework → “Pig latin”

MapReduce Job – Logical View



Nasjonalbiblioteket

Map Reduce is the key processing concept of Hadoop

Controlling processing in the global environment from a single point
Pig script language similar syntax to unix shell scripts
Pig extendable with java programs (JRE/jar) distribution of the jar file

HADOOP handles distribution of both data and processing.

Nasjonalbiblioteket

Other commonly used tools with Hadoop

“Hive” key value database system, Limited sql support.
Lucene → Advanced file based indexing system (from the google index..).
Mahout → decision support framework for implementing reasoning with large data.
AWS → Implementing web services on top.
Monitoring tools

Nasjonalbiblioteket

Language processing and big data; New possibilities

- Language processing is embarissingly easy to distribute
- And to run massively in paralell
- Tokenization is the heart of language processing.
- Different language processing goals requires different tokenization
- Add Machine learning techniques and semantic knowledge.

Language processing and big data with hadoop; perfect match

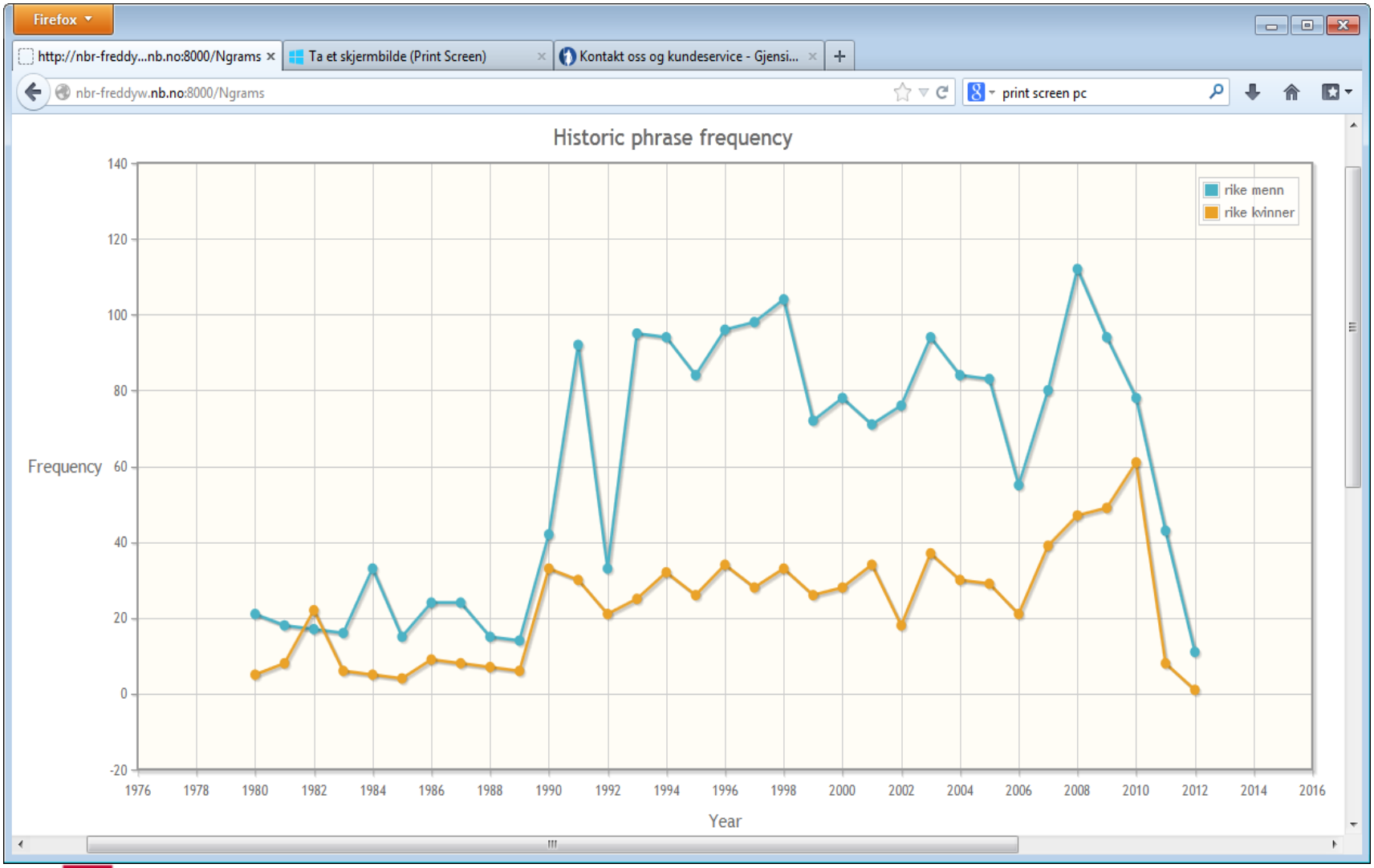
- Large collections of data in a distributed storage
- Different configurable tokenizations (rule based,tree bank,regexp,topic shifts,words,sentences,actions,scenary..)
- Mahout for modeling of semantic knowledge.
- Hugely paralizable processing engine at your fingertips.

Nasjonalbiblioteket

Starting up → N gram

- Tokenization and indexing all text in the National library (newspapers,books, publications).
- Everything we have digitized and will digitize :-)
- Representing metadata and semantic data about the different texts
- Providing instant N gram analysis to all text with references back to text.
- Extending the facilities to open for more relational analysis (Ngrams given...)
- Reasoning with this using Mahout (Apache Reasoning framework).
- Circumstancial processing.

Ngram example in NB texts; Term usage through the years



Nasjonalbiblioteket

NLP loopback

- Using Ngrams to enrich the quality of production lines.
- Improving quality of text with more complex character settings → error correction.
- Building wordlists with different approaches. For instance, a particular time frame, newspaper genre etc.

Nasjonalbiblioteket

Where do we go from here?

- Text based reasoning engine?
- Towards a framework for representing and doing machine extrable knowledge(semantics).
- From text to metadata to knowledge.
- Better understanding of the user.

Nasjonalbiblioteket