Provenance in Big Data

Rajeev Agrawal, North Carolina A & T State University

ncds | THE NATIONAL CONSORTIUM for DATA SCIENCE

Big Data

- Is big data really big data ?
- Big data is a buzzword used to describe the rapid growth of both structured and unstructured data.
- Social networking websites such as Facebook, Twitter, Linked-In, Google+, Youtube etc.

reddit
Google+
YouTube
twitter
NING
flickr
LinkedIn
Meetup
facebook
Pinterest
digg
slideshare
WordPress
delicious
myspace

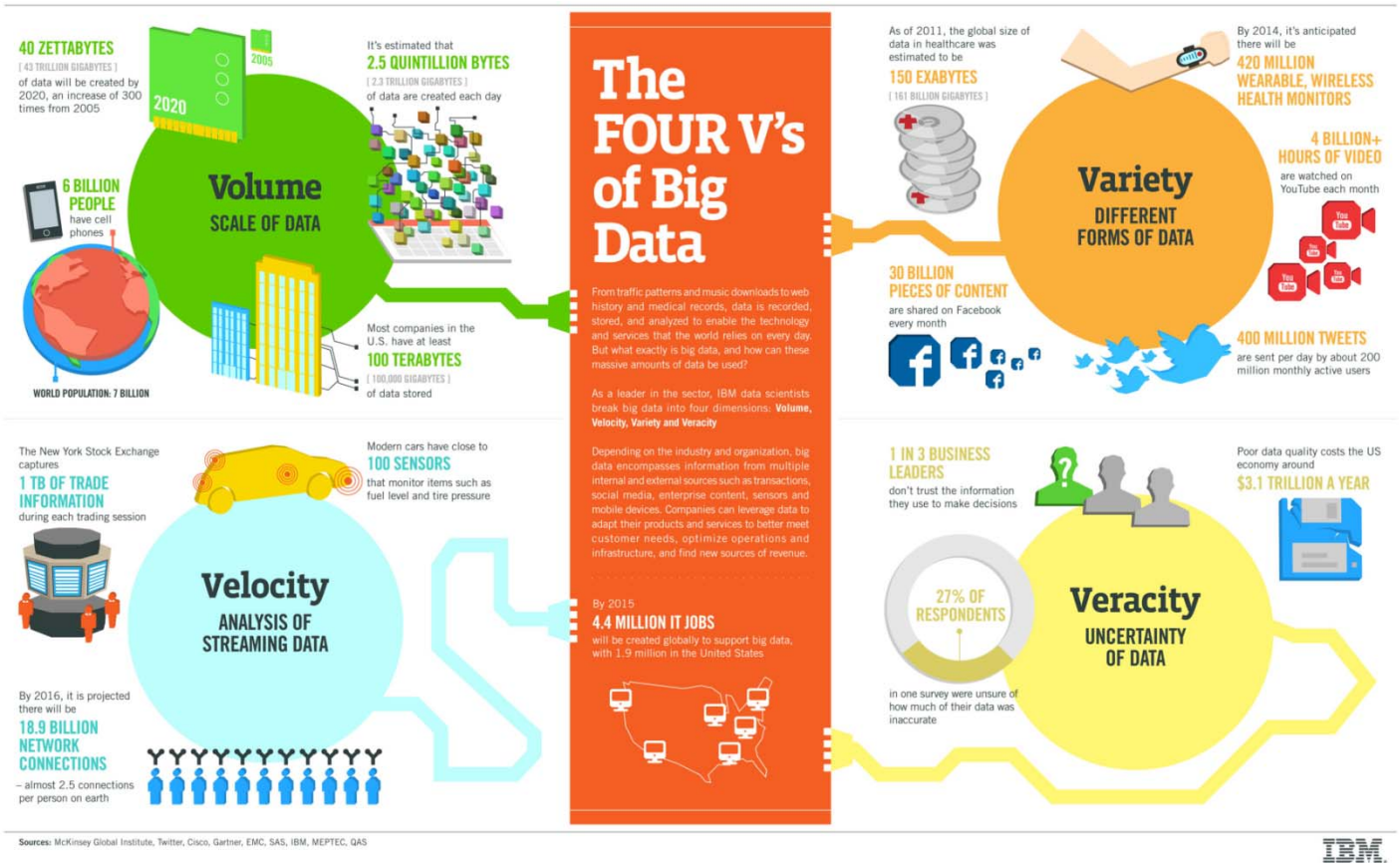
SOCIAL
MEDIA

- “There is a big data revolution”.
- But the revolution is not about the quantity of data rather using the data and doing a lot of things with the data.
- To understand the phenomenon that is big data, it is often described using five Vs: Volume, Velocity, Variety, Value, and Veracity.

Further….

- **New data management systems** that handle a wide variety of data from sensor data to web and social media data.
- **Improved analytical capabilities** (sometimes called advanced or big analytics) including event, predictive and text analytics.
- **Faster hardware** ranging from faster multi-core processors and large memory spaces, to solid-state drives and tiered data storage for handling hot and cold data.

Big Data


40 ZETTABYTES
[43 TRILLION GIGABYTES]
of data will be created by 2020, an increase of 300 times from 2005
2005
2020
It's estimated that
2.5 QUINTILLION BYTES
[2.3 TRILLION GIGABYTES]
of data are created each day
6 BILLION PEOPLE
have cell phones
WORLD POPULATION: 7 BILLION
Volume
SCALE OF DATA
Most companies in the U.S. have at least
100 TERABYTES
[100,000 GIGABYTES]
of data stored
The FOUR V's of Big Data
From traffic patterns and music downloads to web history and medical records, data is recorded, stored, and analyzed to enable the technology and services that the world relies on every day. But what exactly is big data, and how can these massive amounts of data be used?
As a leader in the sector, IBM data scientists break big data into four dimensions: Volume, Velocity, Variety and Veracity
Depending on the industry and organization, big data encompasses information from multiple internal and external sources such as transactions, social media, enterprise content, sensors and mobile devices. Companies can leverage data to adapt their products and services to better meet customer needs, optimize operations and infrastructure, and find new sources of revenue.
By 2015
4.4 MILLION IT JOBS
will be created globally to support big data, with 1.9 million in the United States
As of 2011, the global size of data in healthcare was estimated to be
150 EXABYTES
[161 BILLION GIGABYTES]
By 2014, it's anticipated there will be
420 MILLION WEARABLE, WIRELESS HEALTH MONITORS
4 BILLION+ HOURS OF VIDEO
are watched on YouTube each month
Variety
DIFFERENT FORMS OF DATA
30 BILLION PIECES OF CONTENT
are shared on Facebook every month
400 MILLION TWEETS
are sent per day by about 200 million monthly active users
The New York Stock Exchange captures
1 TB OF TRADE INFORMATION
during each trading session
Modern cars have close to
100 SENSORS
that monitor items such as fuel level and tire pressure
Velocity
ANALYSIS OF STREAMING DATA
By 2016, it is projected there will be
18.9 BILLION NETWORK CONNECTIONS
– almost 2.5 connections per person on earth
1 IN 3 BUSINESS LEADERS
don't trust the information they use to make decisions
Poor data quality costs the US economy around
$3.1 TRILLION A YEAR
27% OF RESPONDENTS
in one survey were unsure of how much of their data was inaccurate
Veracity
UNCERTAINTY OF DATA
Sources: McKinsey Global Institute, Twitter, Cisco, Gartner, EMC, SAS, IBM, MEPTEC, QAS
IBM

Big Data Landscape

Vertical Apps
PREDICTIVE POLICING, bloomreach, MYRRIX

Ad/Media Apps
rocketfuel, collective[i], bluefin, Recorded Future, LuckySort, Media Science, TURN, DataXu

Business Intelligence
ORACLE | Hyperion, SAP Business Objects, RJMetrics, Microsoft Business Intelligence, IBM COGNOS, birst, Autonomy, MicroStrategy, QlikView, bime, Chart.io, GoodData

Analytics and Visualization
tableau, Palantir, OPERA, metaLayer, METAMARKETS, dataspora, centrifuge, TERADATA ASTER, SAS, TIBCO, KARMASPHERE, panopticon, pentaho, Datameer, platfora, ClearStory, CIRRO, alteryx, visual.ly, AYATA

Log Data Apps
splunk, loggly, sumologic

Data As A Service
factual, kaggle, knoema, GNIP, DATASIFT, Windows Azure, INRIX, LexisNexis, LOQATE

Analytics Infrastructure
Hortonworks, VERTICA, MAPR, cloudera, INFOBRIGHT, ParAccel, EMC², GREENPLUM, NETEZZA, kognitio, DATASTAX, EXASOL, calpont

Operational Infrastructure
CouchBase, 10gen, TERADATA, HADAPT, TERRACOTTA, VoltDB, MarkLogic, INFORMATICA

Infrastructure As A Service
amazon web services, Windows Azure, infochimps, Google BigQuery

Structured Databases
ORACLE, MySQL, SQL Server, PostgreSQL, IBM DB2, SYBASE, memsql

Technologies
hadoop, hadoop MapReduce, mahout, APACHE HBASE, Cassandra

Copyright © 2012 Dave Feinleib

dave@vcdave.com

blogs.forbes.com/davefeinleib

What is Actually Big Data?

- In simple words, big data can be defined data that's
  - too big,
  - too quick or
  - too hard for existing tools to process and analyze.
- Here, "too big" means that organizations increasingly must deal with gigabyte-scale or petabyte-scale collections of data.
- For example, *Facebook* generates and stores four images of different *sizes*, which translates to a total of 60 billion images and 1.5PB of *storage*.
- "Too quick" means data is not only huge enough but also it must be processed and analyzed quickly – For example, to identify fraud at a point of sale or transaction.
- Lastly, "too hard" means data may not follow any particular structure. As a result, no existing tool can not process and analyze it properly.
- For example, data that is created in media, such as MP3 audio files, JPEG images and Flash video files, etc.

Provenance

- Data provenance refers to the description of the origin, creation and propagation process of data.
- Data provenance is the lineage and derivation of the data. It stores ownership and process history about data objects.
- Provenance has been studied extensively in the past in arts, literary works, manuscript etc.
- The provenance of data is crucial for
  - validating,
  - debugging,
  - auditing,
  - evaluating the quality of data and
  - determining reliability of data.
- Provenance has been typically considered by the database, workflows and distributed system communities.

Examples of Data Provenance

- Shore Side Data System (SSDS) https://marinemetadata.org/blogs/graybeal/nice-example-of-data-provenance
- Panda http://infolab.stanford.edu/panda/
- Electronic Health Record Data http://repository.academyhealth.org/cgi/viewcontent.cgi?article=1058&context=egems
- A Primer on Provenance http://queue.acm.org/detail.cfm?id=2602651
- Big Data Provenance Examples:
- Lipstick (Putting Lipstick on Pig): http://db.cis.upenn.edu/research/UNIPROVE.html
- RAMP http://infolab.stanford.edu/~hyunjung/ramp/

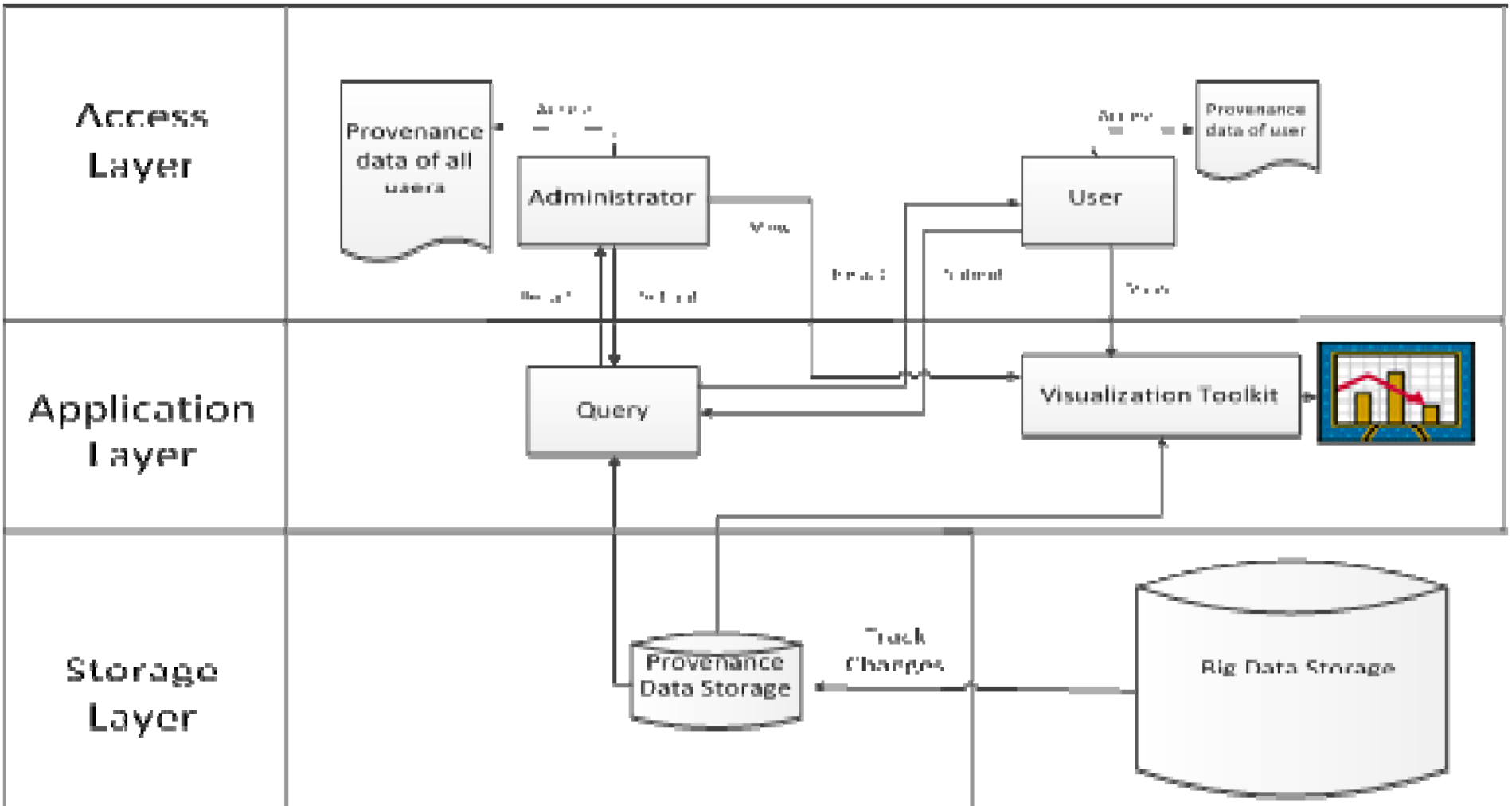
Challenges of provenance in Big data

- Difficult to access all the data
- Difficult to analyze the data
- Difficult to share information and insights with others
- Queries and reports take long time to run
- Expertise needed to run the analysis legitimately

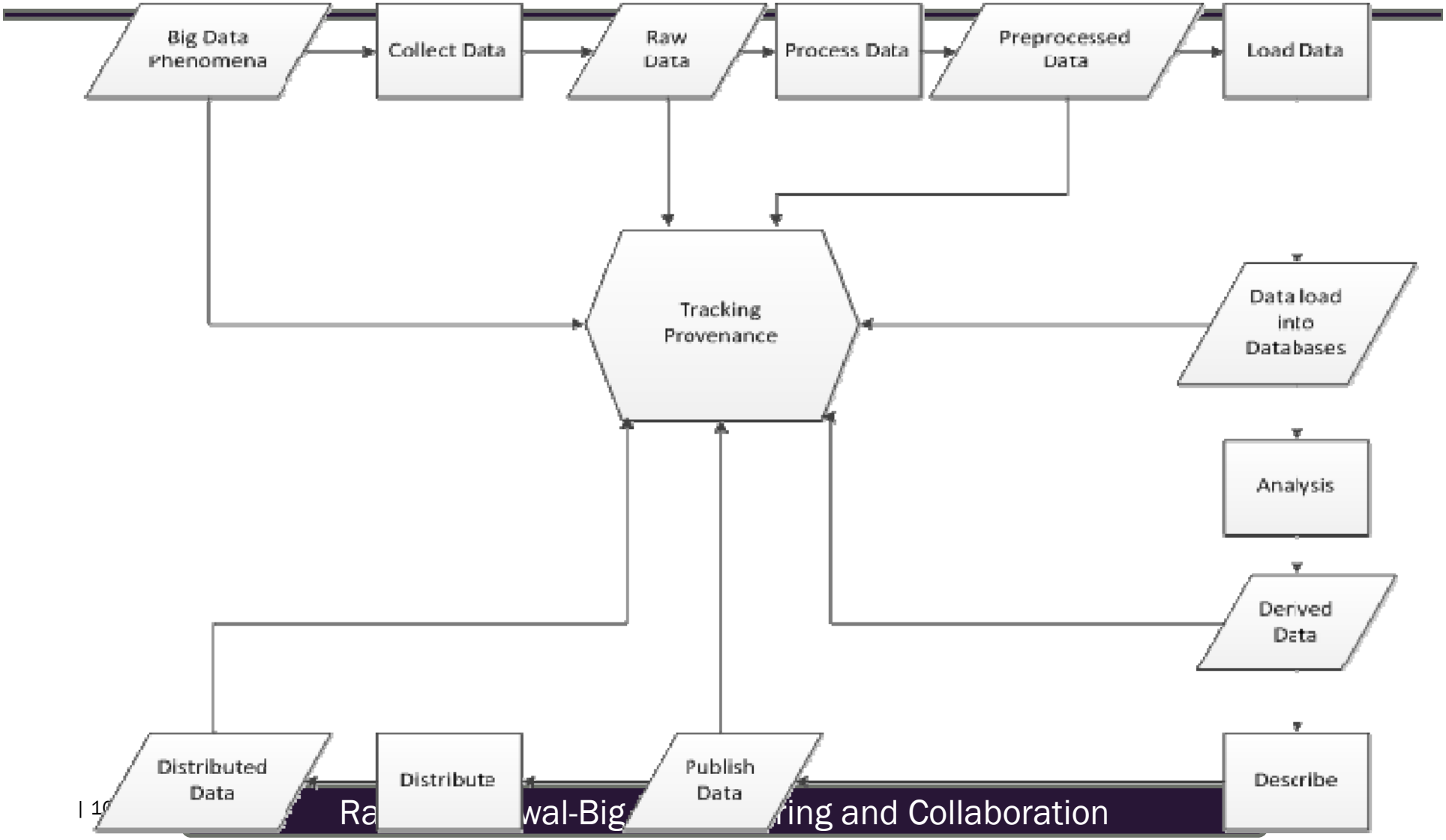
Design goals for a Provenance system

- Less overhead
- User annotations support
- Visualization support
- Scalability
- Various data models support
- Secure Provenance
- Flexibility

Layer based Architecture



Provenance collection



Components

- Storage
  - NoSQL Database
    - MongoDB
- Visualization
  - User based visualization
  - Time based visualization

Data Storage Format in MongoDB

```
{ "_id" : 20020,
"Name" : "Bob",
"Occupation" : "Student",
"Age" : 28,
"Street Address" : "200 Battleground",
"State" : "North Carolina",
"City" : "Greensboro",
"Phone" : 3368730932 }
```

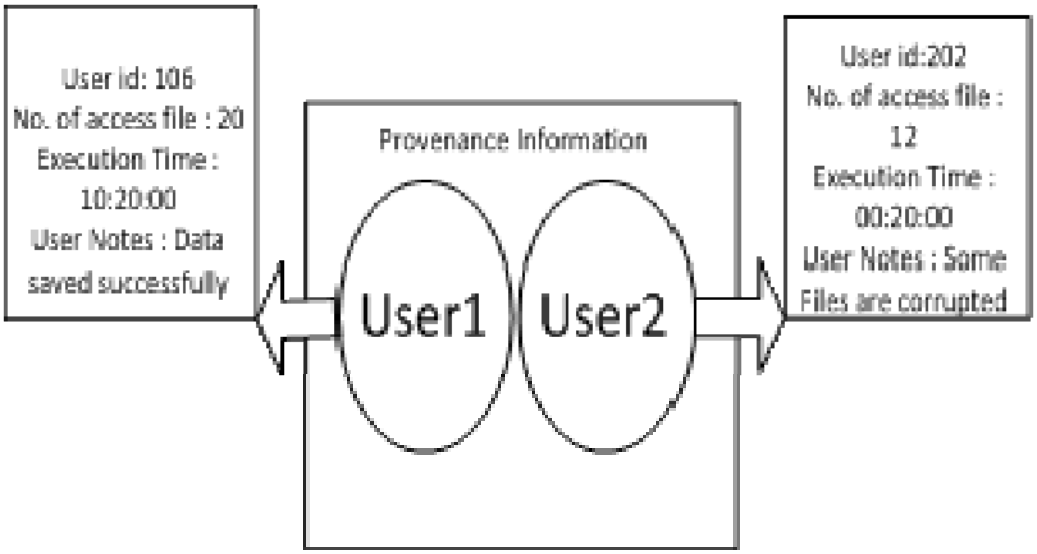
File Management System

- For file management, we use the GridFS specification for storing the actual file contents
- It provides metadata and large file support.
- Useful for storing and retrieving files that exceed the BSON-document size limit of 16MB.
- GridFS divides a file into parts, or chunks,
  - Stores each of those chunks as a separate document.
  - By default, GridFS limits chunk size to 255k.
  - GridFS uses two collections to store files. One collection stores the file chunks, and the other stores file metadata.
- GridFS is useful not only for storing files that exceed 16MB but also for storing any files for which you want access without having to load the entire file into memory.

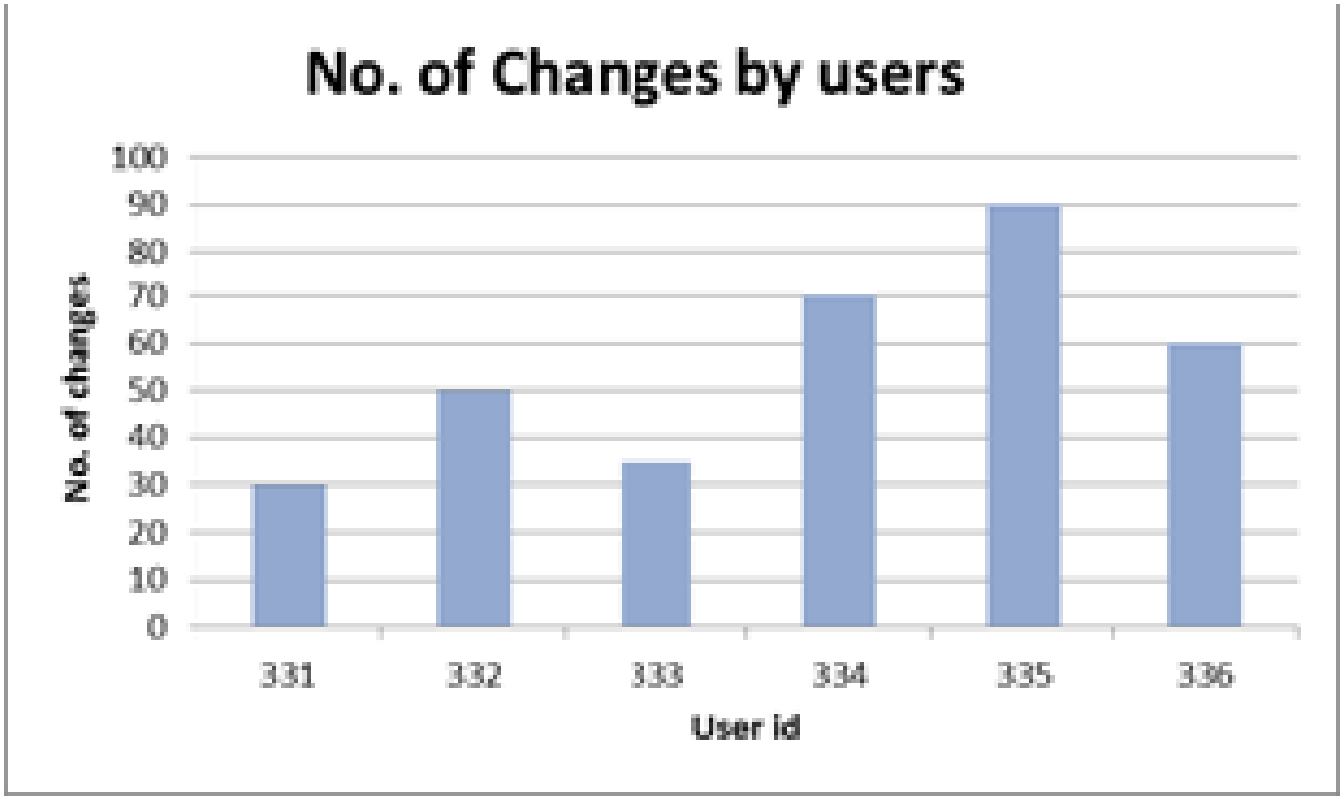
GridFS files collection prototype in MongoDB

```
{ "_id" : <ObjectId>,
"length" : <num>,
"chunkSize" : <num>,
"uploadDate" : <timestamp>,
"md5" : <hash>,
"filename" : <string>,
"contentType" : <string>,
 "aliases" : <string array>,
 "metadata" : <dataObject>, }
```

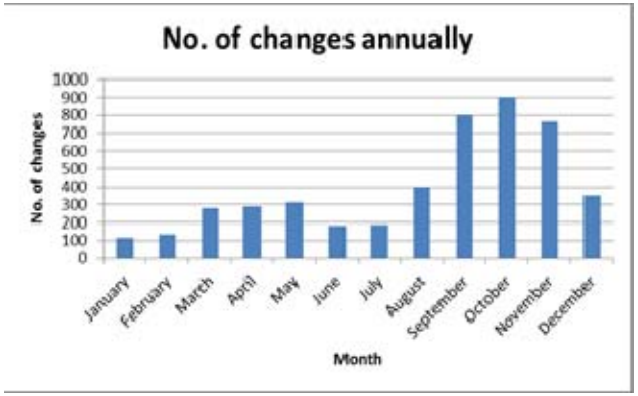
Provenance Information(User based)



Provenance information in a typical day(time based)



Provenance Information in a year by a user(time based)



Conclusions

- Identified design goals
- Layer based architecture
- Access control mechanism to address security
- Handle both structured and unstructured data.
- It can handle simple queries only
- We will add additional component to handle complex queries

- Questions????

THANK YOU