Big Data, Open Data and Research Data Management

Odisha Library Academy Webinar Series- II
23rd May 2020

ARD Prasad
Former Professor
DRTC, ISI

Agenda

- Data
- Big Data
- Open Data
- RDM

Data

Sources of Data

- Data from Labs, Universities, Industries from Research Projects, Surveys, Theses and Dissertations
- Streaming Data from various equipment (satellites, sensors, CCTVs, biomedical) and **social media**

Avatars of Data

- Experimental
- Clinical
- Observational
- Survey

- Numeric
- Textual
- Visual
- Digital or Physical

More Avatars of Data

- **Structured Data** – Tabular data as in the case Relational Data Bases Example: MySQL, postgreSQL

- **Semi-Structured Data** – Data which has some structure but cannot be saved in a tabular form in relational databases is known as semi structured data. Example:XML data, email messages etc.

- **Unstructured Data** – Example- Video files, Audio files, Text file having no structure etc.

Data Majors

- Government Data
- **Research Data**

Big Data

- Big data is huge and **traditional data processing applications are inadequate** Challenges include capture, curation, search, sharing, storage, transfer, analytics, visualization, and information privacy & security.
    - – wikipedia (modified)

Three V's of Big Data

- Volume (Terabytes to Zettabytes)
- Verity (structured & unstructured)
- Velocity (Batch processing to Streaming Data)

    - Veracity (bias, noise and abnormality)
    - Validity (trustworthiness
    - Volatility (current or obsolete)

Volume of Data

- **Facebook** handles **30+ petabytes** of user generated data
- **Youtube** users upload 48 hrs. of video every minute
- **Twitter** gets 175 million tweets everyday (2012)
- **Google** processes 20,000 TB of information a day

Note: Social media data is heavily used for business analytics

Stakeholders of Big Data

- **Data Capture (Information Science people)**
  - Formats & structure of data
  - Standards for interoperability & discovery – metadata & ontologies
  - Filtering & weeding out irrelevant data
  - Data Curation: ensuring long term preservation and reuse
- **Technology for Big Data (Technology people)**
  - Hadoop Eco System, NoSQL etc.
- **Data Analytics (Statisticians)**
  - Descriptive, Predictive, & Prescriptive
- **Domain Experts**

Vertical Vs. Horizontal Scaling

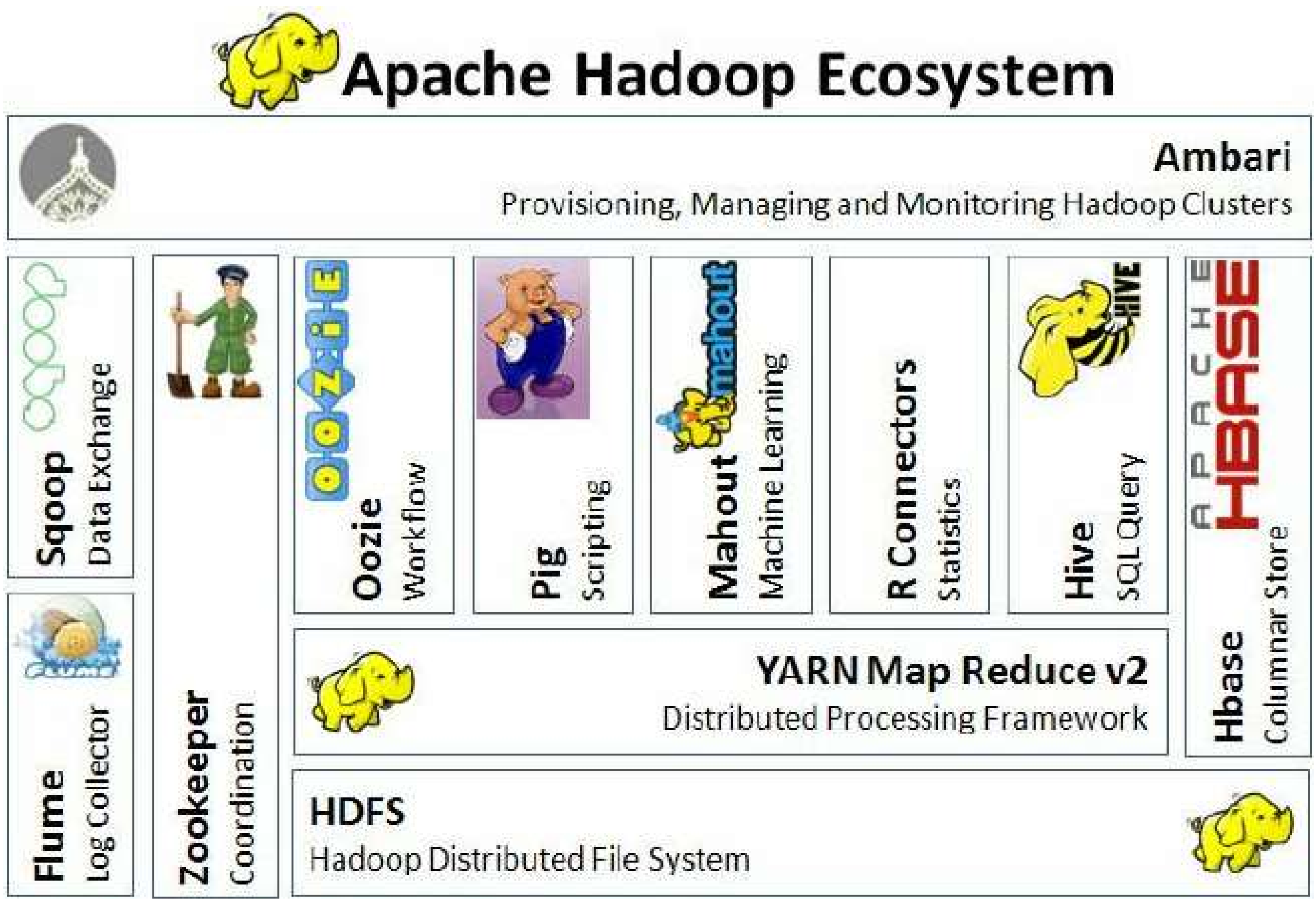
- **Vertical scaling** means that you scale by adding more power (CPU, RAM) to your existing machine
- **Horizontal scaling** means that you scale by adding more machines into your pool of resources

In Data Bases

A good example for horizontal scaling is Cassandra , MongoDB. Vertical scaling is MySQL

NoSQL DBMS (Non RDBMS)

- Key / Value Based
  - Redis, MemcacheDB, etc.
- Column Based
  - Cassandra, HBase, etc.
- Document Based
  - MongoDB, CouchDB, etc.
- Graph Based
  - AllegroGraph, Neo4J, etc.


Apache Hadoop Ecosystem
Ambari
Provisioning, Managing and Monitoring Hadoop Clusters
Sqoop
Data Exchange
Flume
Log Collector
Zookeeper
Coordination
Oozie
Workflow
Pig
Scripting
Mahout
Machine Learning
R Connectors
Statistics
Hive
SQL Query
Hbase
Columnar Store
YARN Map Reduce v2
Distributed Processing Framework
HDFS
Hadoop Distributed File System

Analytics and Visualisation

Open Source Software: R, R Studio

Commercial: MatLab

Predictive Analytics

- It utilizes a variety of **statistical, modelling, data mining, and machine learning techniques**
- Predictive analytics can only **forecast** what might (NOT WILL) happen in the future, because such analytics are **probabilistic in nature**.

Prescriptive Analytics

- Recommend one or more **courses of action** -- and showing the likely outcome of each decision.

- Source: Dr. Michael Wu

Open Data

Re-usability

- It is impossible to imagine the **evolution of mankind** without reusing the experience, knowledge of the past generations
- Evolution of **languages** ensured effective communication of the acquired knowledge to the successive generations
- **Writing(scripts)** enhanced reusability by recording knowledge
- **Libraries played an organised/institutionalised role** in reusability by preserving and disseminating knowledge

Worldwide Movement
Open Mantra

- Open Source Software
- Open Access to Information/Content
- Open Standards
- Open Data Repositories
- Open Science/ Open Research

Open Notebook Science

- Practice of making the entire primary record of a research project publicly available
- Failed, less significant, partial and otherwise unpublished experiments; so called 'Dark Data'
- In case of Governments too have unpublished Dark Data

Why data is not published?
Many Publications used data

- Publishers point of view
  - For lack of space (not in case of Web)
  - Not quite profitable

    Recently some publisher are insisting

- Author point of view
  - Author might have overlooked the data
  - Author deliberately did not present data so that others can not verify the data

Example

Some suspect that Sigmund Freud's data is of fictitious persons, not just fictitious names

–

Controversy that some particle is moving faster than light

–

Aaron Swatch (JSTOR)

Closed Data (not completely open)

- Data is not published
- Subscription based access
- Access to registered users
- Encrypted data
- Data requires proprietary tool to access
- Copyright/license/patent forbidding reuse
- Not allowing robots/spiders, CAPTCHA to access data
- Time-limited access, not allowing bulk downloads
- Political, legal, commercial pressure on restricting or banning access – (Boris Pasternak, Salman Rushdie)

In Support of Open Data

- Data belongs to Mankind
- Mostly data is generated by Public Money
- Facts can not be copyrighted
- Data value will be fully realised if it is widely used, reused
- Restrictions will result in anti-creative-commons
- Open data will create more harmony
- Will accelerate more scientific research

If data is openly available …

- Others may draw different conclusions, sometimes, contradictory to that of the author
- Others may deal with other facets of the data
- Data Transparency supplements the Objectivity and self corrective characteristics of Science

- Note: If “Case history of patients” is openly available, it will contribute significantly to medical research

Information/Digital Divide

- Open Access Journals and Institutional Digital Repositories helped bridging gap in digital divide to a large extent, especially in Humanities and to a lesser extent in Social Sciences and even lesser in Physical and Natural Sciences
- Physical and Natural Sciences do require laboratory infrastructure

Against Open Data

- Data generated by public money will be used by private organisations
- Privacy concerns
- Data collected, cured by private organisations should get back their investment
- Data was collected using costly equipment or hired manpower

Philosophy

- Data should be freely available with out restrictions such as
  - Biased copyright laws,
  - Some ridiculous patents etc.
- Philosophy and Sociology of Science should guide us

Bad Side of Science

- Science is vastly used for
  - defence purposes (to kill people)
  - profit making (to rob people)
  - Of course, not without a few good side effects

Debatable

- Data in wrong hands
- How to make crude Bombs ?
- Some issues related to pornography
- Governments Vs. Terrorism:
  - Terrorists misuse information
  - In the name of anti-terrorism Governments encroach into privacy of people
  - George Orwell was short-sighted

Research Data Management

What is RDM?

The term **Research data management** includes organising, structuring, storing the data generated during a research project

Covers every phase of Research Data Life Cycle

Why RDM?

- Increases individual and **institutional reputation** because the **data can be cited**
- Improves the **quality of research** by ensuring data validation
- Reduces **duplication** of research
- **Avoids loss** of data
- **Streamlines research process**

Why RDM?

- **Funders** can know how the data is being used and support research projects
- Some funding agencies already mandated **Data Managemnt Plan (DMP)** in the proposal

RDM Rules

- Understand how institutions deal with research data
- Institution's take on RDM to establish policy and strategy
- Ensure researchers are aware of what data is available
- Provide easy to use, robust data storage
- Make it easy for others to find and cite research data

Data Mangement Plan (DMP)

- Is a formal document that outlines how data are to be handled both during a research project, and after the project is completed -- Wikipedia
- A requirement by many funding agencies
- Comply with legal and ethical guidelines

- **NOTE:** We developed an online model to be adopted by Indian funding agencies

Policy Should be the Guide

- Governement should mandate
  - National Data Sharing and Accessibility Policy (NDSAP) 2012
- Funding Agencies
  - Should insist DMP – NSF, Welcome Trust etc.
- Organisations – Research Labs, Universities
  - Evolve policy in tune with Govt. And Funding agency requirements
  - Digital Curation Centres (DCC) to be established
  - Librarians should be trained

Research Data Life Cycle

- Capture
  - Collect data from Surveys, Experiments from equipments, instruments etc.
  - Structuring: Database tables
  - Formats: Using Open Standards formats
- Cataloguing: Creating metadata, assigning subject descriptors, Linked Open Data, DOI
- Data Repository: Making data discoverable, shareable, allowing harvesting, and reusable

RDM Includes

- Creating/Acquiring Data, Anonymisation
- Converting Data into Open Standard Format
- Adding Metadata
- Classifing Data (Ontologies)
- Adding Licensing
- Adding Persistent Identifiers
- Hosing Data Repositories
- Backup

A Few Examples of Research Data

- Maps
- Genome Data
- chemical compounds
- Bio-medical data and case histories
- Government Data
- GIS data
- Weather Data
- Simulation Data
- Log Data
- Social Media data
- Survey data

FAIR Data

- Findable
- Accessible
- Interoperable
- Reusable

5 Star Data
Tim Berners-Lee

- Make ...
  - data available on the Web Under open license
  - data available as structured data
  - data available in a non-proprietary open format (e.g., CSV instead of Excel)
  - use URIs to denote things, persisten Id
  - link your data to other data to provide context (LOD)

Open Archival Information System
OAIS Reference Model

- Purpose
  - To build trusted repositories
  - Facilitate analysis and comparison of repositories
  - Informing system design
  - Preservation metadata

Role of libraries in RDM

- ***Providing access to data:*** Traditional library services include consultation and reference service for researchers looking for data
- ***Awareness and support for managing data:*** Educating researchers about the importance of data management and hands on support for data management life cycle
- ***Managing a data collection :*** This includes data collection, data management , data preservation – DATA CURATON

Core Competencies for librarians

- Some level of subject knowledge in order to understand the domain properly
- How to provide access to data centres, repositories etc.
- Knowledge of policy and standards for RDM
- Knowledge of Data management tools
- Knowledge of metadata schema, data formats, domain ontologies

Core Competencies cont...

- Linked Open Data (**LOD**)
  - Linking and data integration techniques
- **Data repositories** and storage platforms
- **Data citation** and referencing practices
- **Research practices** and workflows

Points to Ponder

- **Domain specific data** curation strategies
- **No one-size-fits-all solutions**, but alignment ultimately needed
- **Are there common** collection, representation, and service principles?
- What are the **data intensive domains**

File Formats

- Open Vs. Proprietary Formats
- Compressed Vs. Uncompressed
- Open and Lossless Formats
  - RTF
  - XML
  - Uncmpressed TIFF

Some Domain Specific Metadata Schema

- Dublin Core: Considered as the Lowest Common Denominator among metadata Schema
- Darwin Core: Biological Diversity Data
- Data Documentation Initiative (DDI): Social and Behavioural sciences
- Directory Interchange Format: Earth Sciences
- ISO 19115:2003: Geographic data such as maps and charts
- PBCore: Media assets like individual clips and full, edited, aired productions
- Science Data Literacy Project: Astronomy, Biology, Ecology and Oceanography
- VRACore: Visual Objects like images

Ontologies

- To make relations NT, BT, RT explicit
- Assign Unique IDs to each term
- Helps in Discovery and Linked Data
- Existing Thesauri: Inspec, NASA, Agrovoc etc.
- Use OWL or SKOS

Some Open Data Licenses

- Open Data Commons Public Domain Dedication and Licence (PDDL)
- Dedicate to the Public Domain (all rights waived)
- Open Data Commons Attribution License
- Attribution for data(bases)
- Open Data Commons Open Database License (OdbL)
- Attribution-Share Alike for data(bases)
- **Creative Commons CCZero**

Data Citation

We have been providing reference to a research publication. Similarly, we are expected to provide a reference to a data set

Distinct from metadata, as metadata may have much more information like administrative, structural, preservation etc.

Advantages of Data Citation

- Helps in linking the data set(s) to the publication(s)
- A citation study will ensure what are all the publications that have used a particular data set (Impact of the data set – Impact can not be citation alone, there might other tangible and intangible impacts)
- Researcher can find out how others have used the same data set from a different perspective

Typical Citation comprises of...

- Author/Principal Investigator – ARCID
- Title of Data of the Data Set
- Version Number in case the data is updated
- File Format of the Data
- Location of the data set (ex: data repository)
- URI: DOI, CNRI Handles
- Place of Publication
- Year of Publication etc.

Data Validation Software

- Open Refine
- R
- Spark
- Python
- Many Commercial Tools

Extract, Transform and Load

- ETL is a type of data integration that refers to the three steps (extract, transform, load) used to blend data from multiple sources.
  - -- www.sas.com
- S/w Tools
  - Talend
  - Scriptella
  - Pentaho Data Integrator – Kettle
  - GeoKettle

Data Repository Software

- DataVerse
- CKAN
- DKAN
- Dryad
- GeoNode/GeoServer

Registry of Data Repositories

re3data.org

- Currently has 634 entries of data repositories from different disciplines
- Each entry includes
  - URL of the repository
  - A short Description of the repository

Data Search Sites

- Google https://datasetsearch.research.google.com
- ELSEVIER
- https://datasearch.elsevier.com

ISI-DRTC Projects

- **Living Knowledge** (Funded by European Commission) on Semantic Web
- **ITPAR:** India-Trento Program for Advanced Research (Extending PMEST/DEPA to DERA using description logics)
- **AgInfra** (Funded by European Commission): Dealing with Agricultural Data Infrastructure

DRTC/ISI Workshops

- Conducted 2-week international workshop with ICSU/CODATA in March, 2015
- TAB member of RDA, Co-chairing session on Agricultural Data
- Conducted an International Conference on 'Big Data and Knowledge Discovery' (March, 2016)

Thank You

ard@drtc.isibang.ac.in
ardprasad@gmail.com