BIG DATA IN THE CLOUD : CHALLENGES AND OPPORTUNITIES

MARY-JANE SULE
&
PROF. MAOZHEN LI
BRUNEL UNIVERSITY, LONDON

Overview

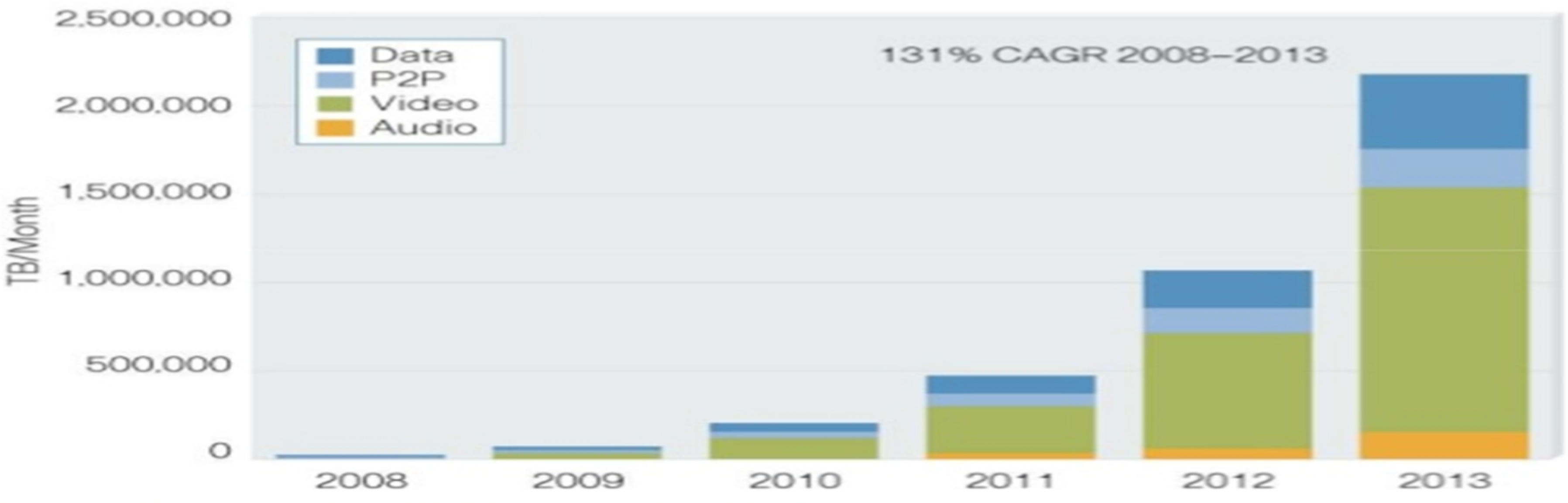
- Introduction
- Multiple faces of Big Data
- Challenges of Big Data
- Cloud Computing : Challenges and Opportunities.
- Suggestions
- Conclusion

Research Area

- Presently there are 5 PhD students and 13 awarded PhDs that are working or have worked in the following research areas together with Prof. Maozhen Li at the Brunel University
- Research topics include :
  - High performance computing (grid and cloud computing)
  - Big data analysis (MapReduce/Hadoop)
  - Semantic Web, artificial intelligence, multi-agent systems
  - Information retrieval, content based image annotation and retrieval
  - Distributed machine learning techniques
  - Context aware mobile computing

INTRODUCTION

- Big Data has always been around, its been relative to its time.



Cisco Forecasts 2 Exabytes per Month of Mobile Data Traffic in 2013

Source: Cisco, 2009

Multiple faces of Big Data

- For scientists, Big-Data is about getting better output from simulations or handling output from simulations
- Data is also coming from sensors (New world of IPV6 is about the Internet of objects) here data can be used by different fields

Integrating Data can never be over –emphasized

CHALLENGES OF BIG DATA

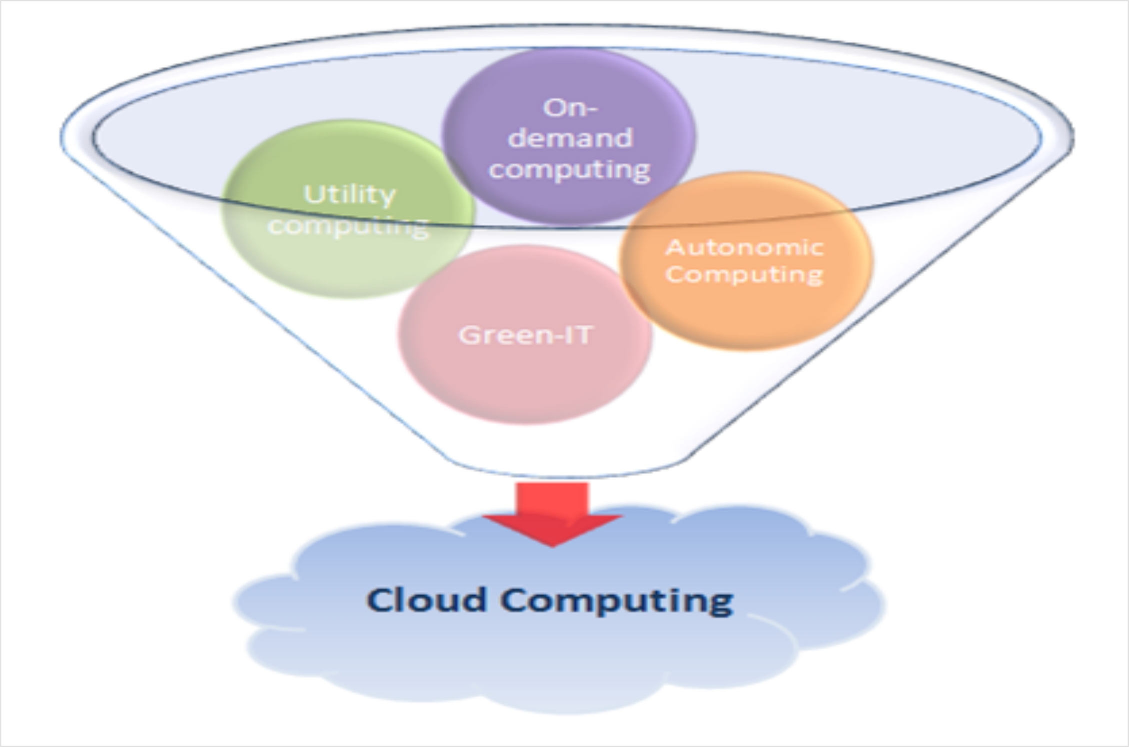
- Storage
  - Clearly not enough hard disks/devices. Distributed storage is still not enough , manufacturers cannot make enough storage devices in time. Speed in writing to devices, bigger data paths/data-bus
- Processing
  - Integrating data using Filters
  - "What" Data and "How" ?
  - Effective Data processing system Design
  - ??Power
  - Internationalisation / Standardisation
  - Latency and Bandwidth
- Taxonomy and Ontology
  - How to classify big data – No standard way of doing that yet
- Security / Privacy
  - What is to be secured – the Data sources ? As IT logs are also now a source of big data

BIG DATA – SQL vs NoSQL

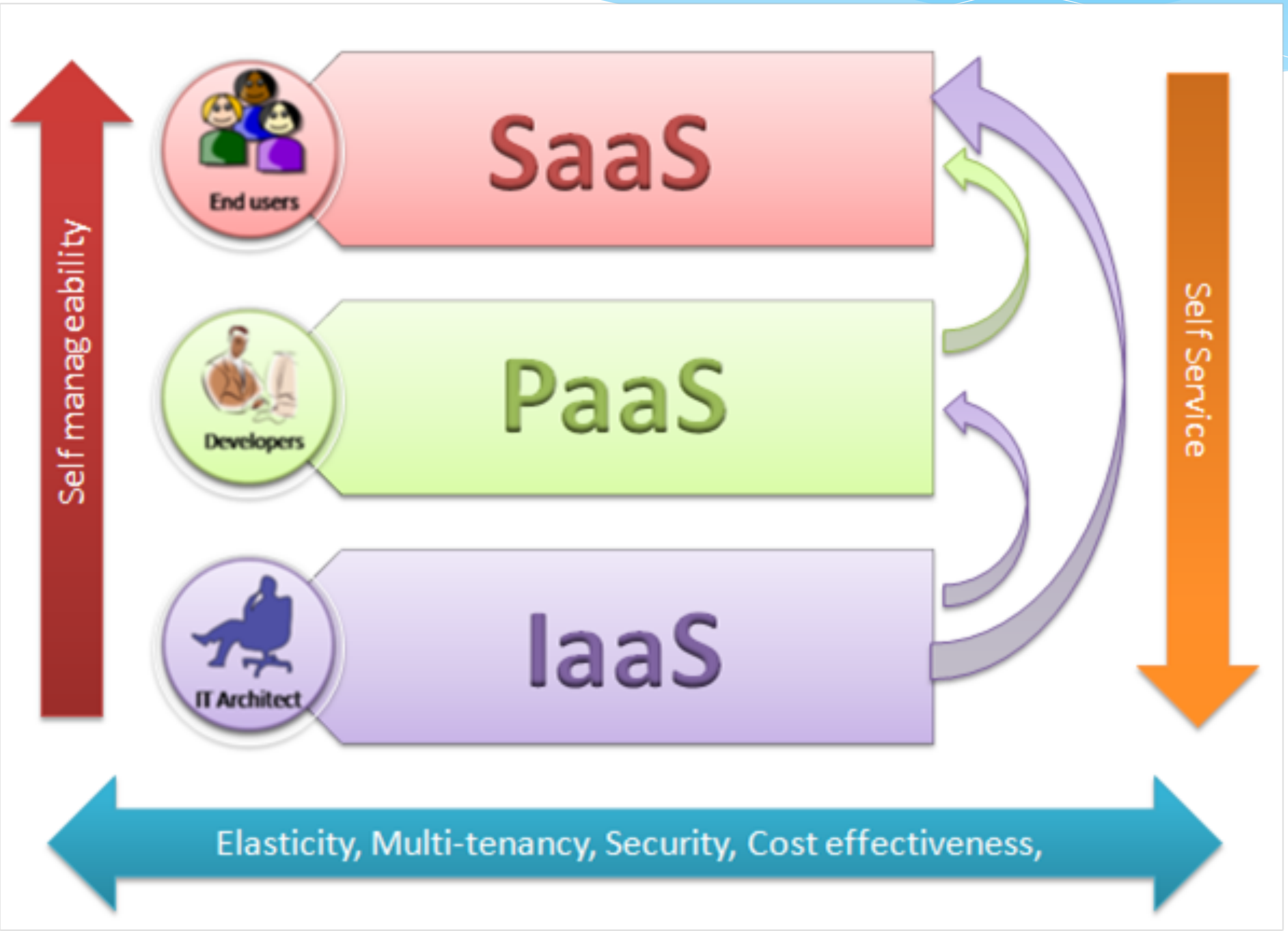
- In Big Data SQL which are used for RDBMS provide ACID
  - Atomicity = all or nothing
  - Consistency = same data values b4 and afta.
  - Isolation = hidden events during transactions (a form of security)
  - Durability = survive subsequent malfunction
- NoSQL provide BASE
  - Basically Available = since it allows parts of the system to fail.
  - Soft Sate = objects with simultaneous values.
  - Eventually consistent = achieved over time.

Cloud Computing

- Definition



Cloud Delivery Types



Cloud Computing

- With Cloud computing here IT expertise is narrowed and professionalism is brought into play. one concentrates on the main issue at hand and builds professionalism and expertise, well grounded in a particular field.
- Allows for horizontal provisioning as against the present vertical provision.(Elasticity)

Cloud Computing and Data Management

- Types of Data
  - Transactional Data Management
    - ACID (Atomicity, Consistency, Isolation, and Durability)
    - Risk / Privacy (especially when placed in clouds)
    - Typically cannot scale on the cloud
  - Analytical Data Management
    - Scalability (Shared-Nothing Architecture)
    - Availability is key over ACID
  - CAP's theorem??

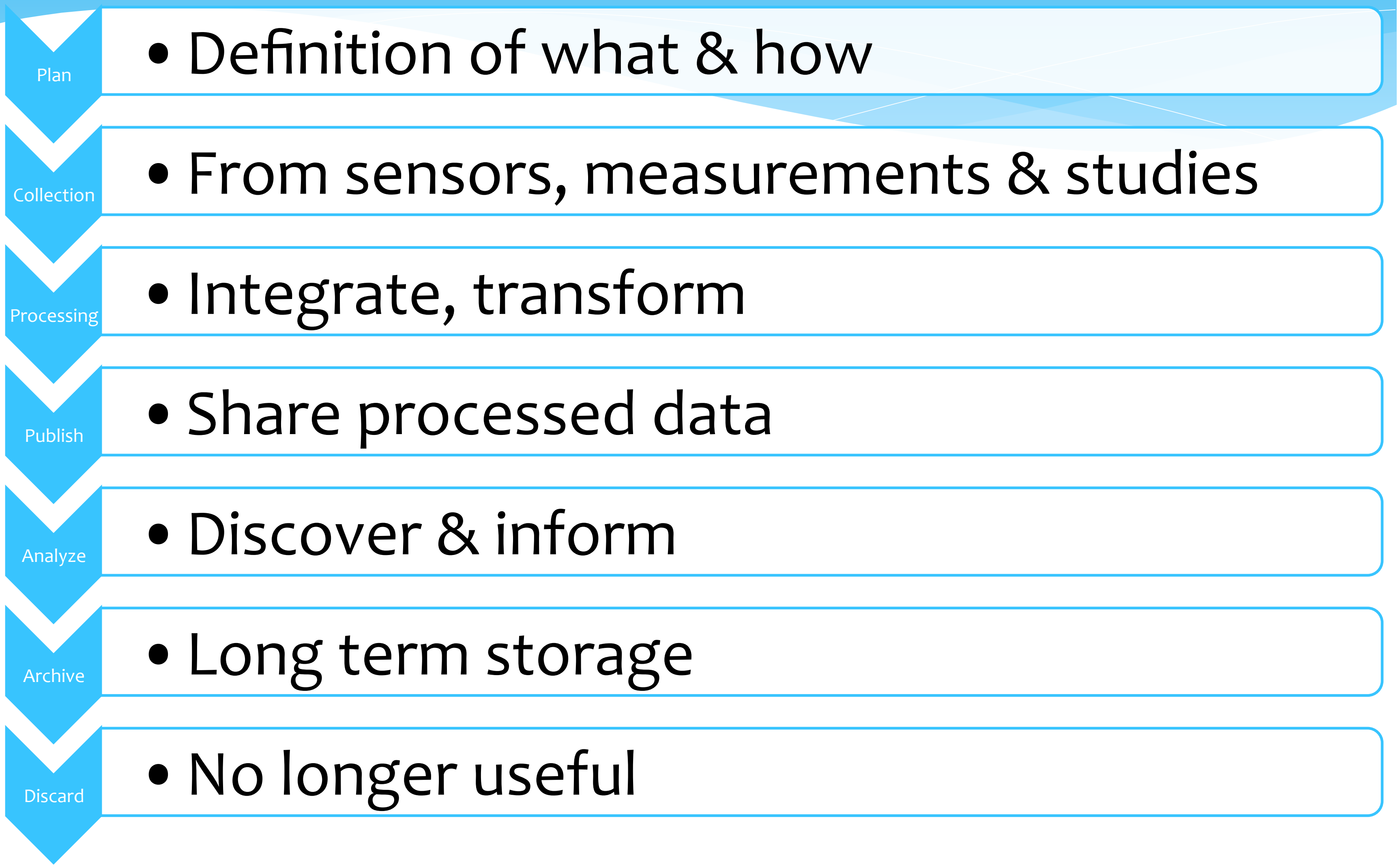
Cloud Computing - Open issues

- The present Cloud computing design has some issues yet to be solved:
  - Basic DBMS has not been tailored for Cloud Computing
  - Data Acts is a serious issue so it would be ideal to have Data Centres located closer the user than the provider.
  - Data Replication must be carefully done else it affects data integrity and gives an error prone analysis
  - Trust in the event of mission critical data.
  - Some deployment models are still in their formative stage.

SUGGESTIONS

- Data management
  - Life-cycle management can start now, once information has been extracted from the data
- Better software design
  - Build software applications around the expected data/information processing and not around CPU processing.
- Cloud computing: Data processing in the could should provide the following :
  - Efficiency
  - Fault tolerance: does not have to restart a query if 1 of d nodes in the query fails but then comes with a trade-off for performance.
  - Ability to run in a heterogeneous environment: shared task across cloud compute nodes.
  - Ability to operate on encrypted data: to build trust data maybe encrypted before uploaded onto the clod and the DBMS application should be able to work on the data without decrypting it.
  - Ability to interface with business intelligence products or other applications

Example of Data Management lifecycle



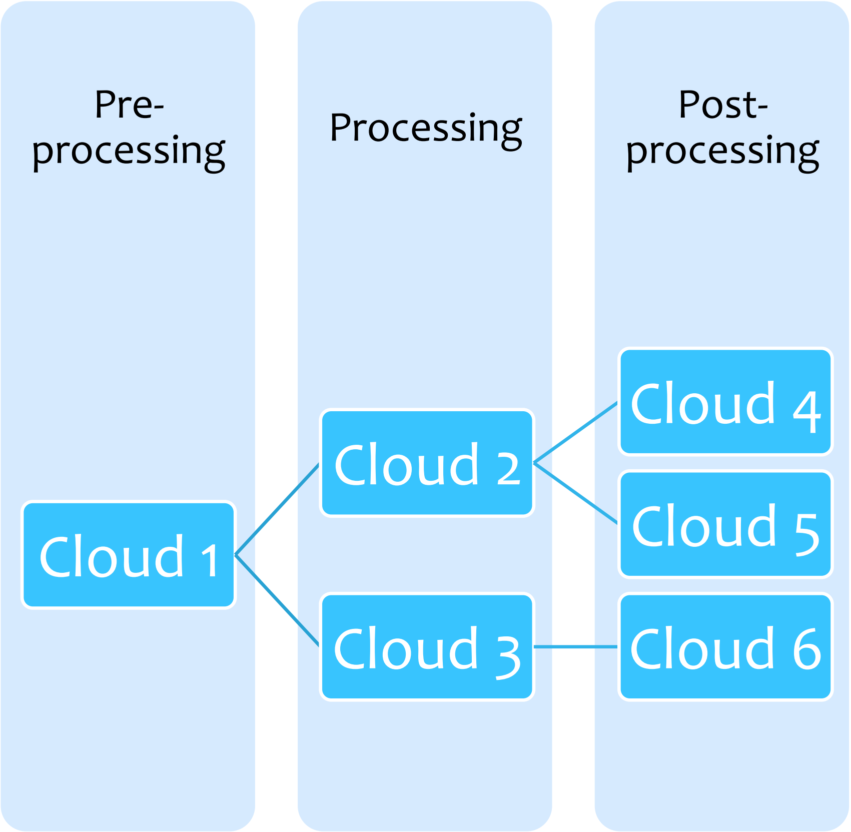
Plan	• Definition of what & how
Collection	• From sensors, measurements & studies
Processing	• Integrate, transform
Publish	• Share processed data
Analyze	• Discover & inform
Archive	• Long term storage
Discard	• No longer useful

The Cloud Opportunities

- Cloud provide opportunities for real time data streams analysis because cloud systems tend to provide scalability - "scale-free", fault tolerance, cost effectiveness and ease of use.
- Current Research areas
  - Science, acquisition, conditioning & evaluation
    - Online Analytics of interest here
  - Framework:
    - similarities to existing frameworks NoSQL & RDBMS
  - Infrastructure
    - How to implement framework(s)

(Note: security & privacy cut across all 3 items)

Suggested Cloud-based data processing model for Scientific applications



Clouds

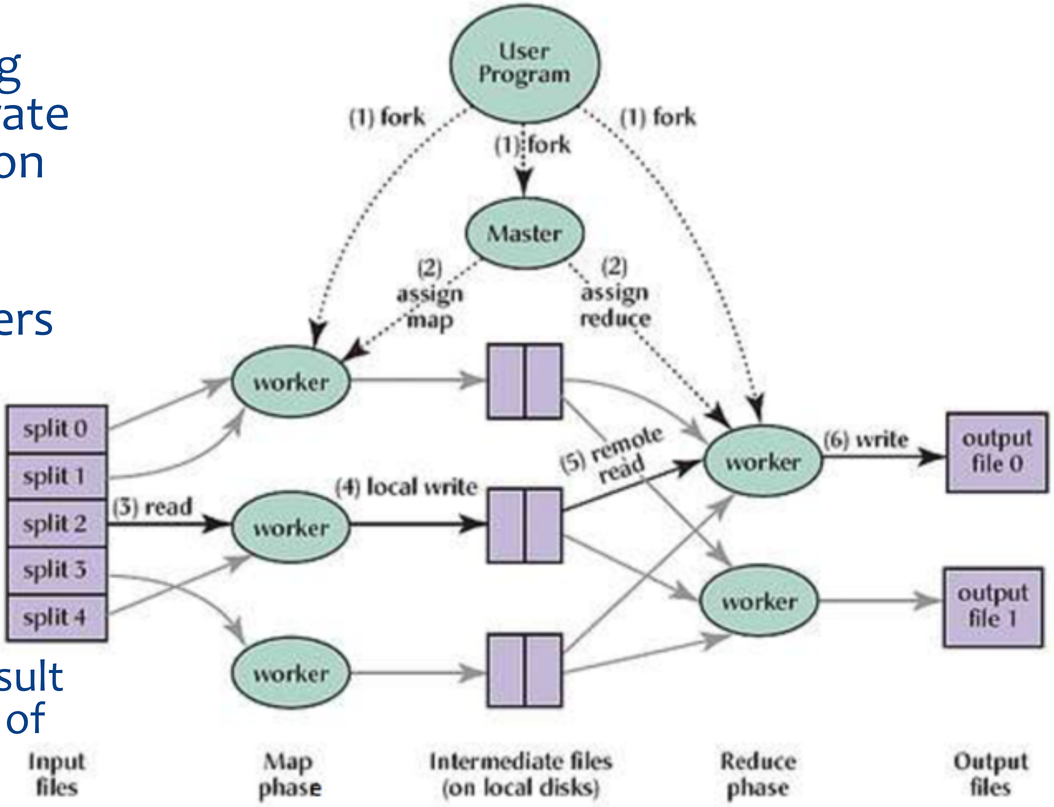
- Scalable sets of cloud arranged around the various data handling activities
- Clouds may be private, public or commercial based on security considerations
- Flexibly add , remove or replace components from pipe-line. E.g cloud 6 provides only visualization of raw output, while cloud 5 may transform output before visualization

Online Analytics

- In time pass, most data were structured and mostly relational, mining for decision making was fairly easy with SQL-like technologies but Big Data has evolved now to include large volumes of unstructured data – speeches, emails, tweets, chats, etc. There came the need for a system that can accept ,process, store, analyze / re-analyze unstructured data to provide value for decision making.

MapReduce
Framework in the Cloud

- MapReduce: is a programming paradigm based on two separate and distinct tasks performed on big data.
- Its allowed for massive scalability across various servers but didn't allow for "plug-ability".
- These separate tasks are:
  - Map : where the data is converted into individual elements and
  - and Reduce : combines the result from the map into smaller set of tuples for further analysis.

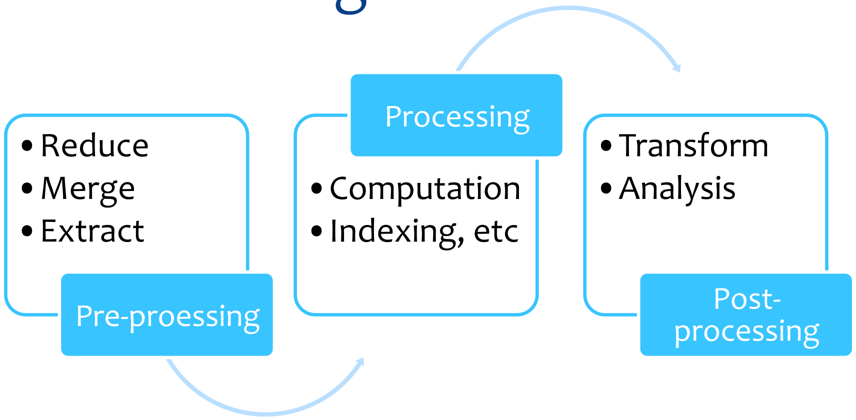


Hadoop: an implementation of MapReduce Framework

- Hadoop is based on the MapReduce framework
- Schedules, monitors and re-executes tasks (this provides for fault tolerance).
- Hadoop stack now includes components for query & storage in addition to MapReduce

An Example

Working with Data files



Batch organisation

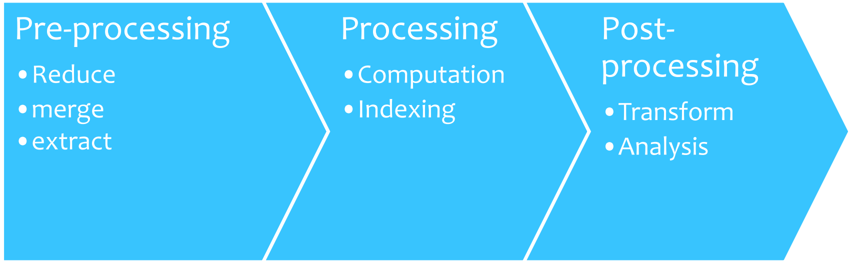
- Files are used as main vehicle for transfer of data between phases and different clouds
- The processing pipe-line is not time dependent and so each item could be implemented as batch operations.
- Example could be HADOOP framework and starcluster (HPC) for processing

MOA as a framework for online analytics

- A framework for Big Data Stream Mining which is written in Java it provides for:
  - Storage for repeatable experiments
  - Set of existing algorithms to use for analysis and
  - algorithms to easily extend the framework for new streams and analysis.
- MOA=>data feed +algorithm + evaluation method = Results
- With MOA data stream mining is done in real time, large scale machine learning
- MOA allows developers to easily extend all the parts of the framework /architecture.
- MOA is also easily used with Hadoop, S4 or storm this provides for a more robust and configurable system

An Example

On-line



Pre-processing

- Reduce
- merge
- extract

Processing

- Computation
- Indexing

Post-processing

- Transform
- Analysis

Dynamic data streaming

- Data is passed from one cloud to another in-line (via network)
- Each cloud in the pipe-line process is dependent on previous phase/cloud.
- Examples: STORM framework and starcluster clouds.

CONCLUSION

- Data processing on a cloud based cluster would provide added benefits such as fault tolerant, heterogeneous, ease of use, free and open, efficient, provide performance and “tool plug-ability” which most DBMS do not provide.
- Combining different types of software such as MOA and Hadoop is a possible solution for online analytics of scientific data. A lot of extension work is still being done with the algorithm to provide even change detection and frequent pattern mining among others.

References & more information

- http://moa.cs.waikato.ac.nz/overview/
- http://www-01.ibm.com/software/data/infosphere/hadoop/mapreduce/

Thank you

:-)