University of Reading

A Novel Cloud Based Elastic Framework for Big Data Preprocessing

Omer Dawelbeit and Rachel McCrindle

Overview

University of Reading

- Introduction
- Cloud based elastic framework
- Motivation
- Major Components
- Workload distribution
- Processing steps
- Experiments and results
- Discussion
- Conclusion and future work

Introduction

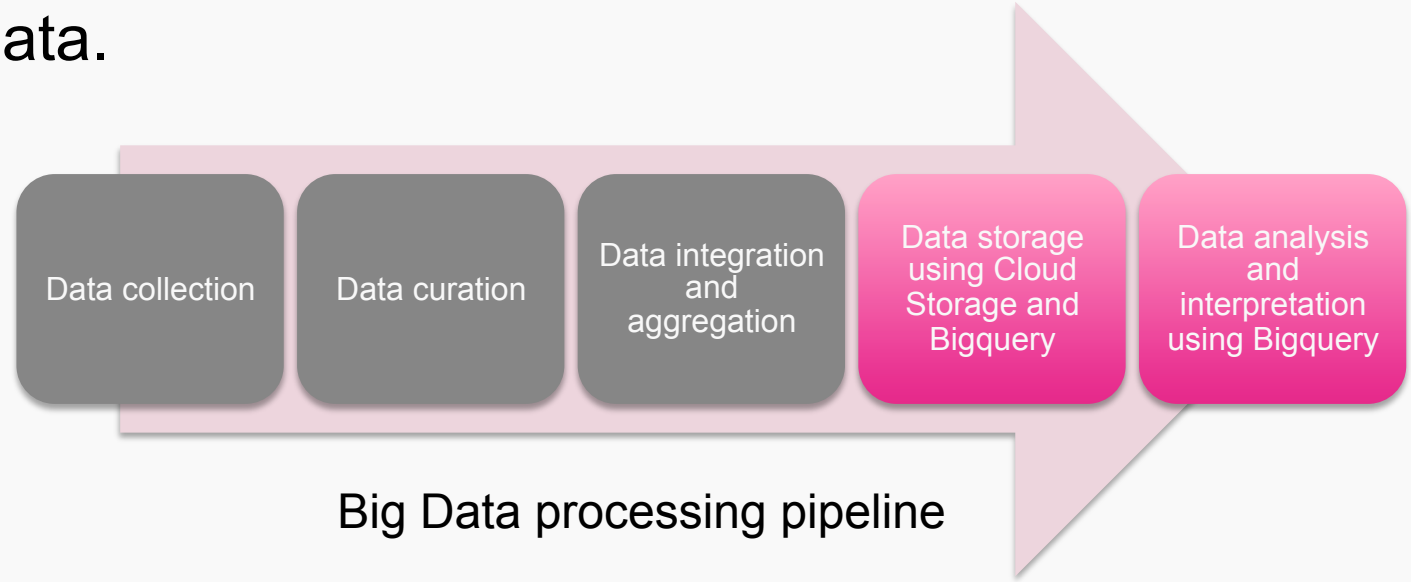
University of Reading

- Big Data is data that is too big, too fast or too hard to process using traditional tools.
- The Primary aspects of Big Data are characterized in terms of three dimensions (Volume, Variety and Velocity).
- Cloud computing is an emerging paradigm which offers resource Elasticity and Utility Billing.
- Cloud computing resources include: VMs, cloud storage and interactive analytical big data services (e.g. Google Bigquery).

Cloud Based Elastic Framework

University of Reading

- Entirely based on cloud computing.
- Elastic, hence able to dynamically scale up or down.
- Extendible, such that tasks can be added or removed.
- Tracks the overall cost incurred by the processing activities.
- Capable of both preprocessing and analyzing Big Data.



Big Data processing pipeline

Motivation

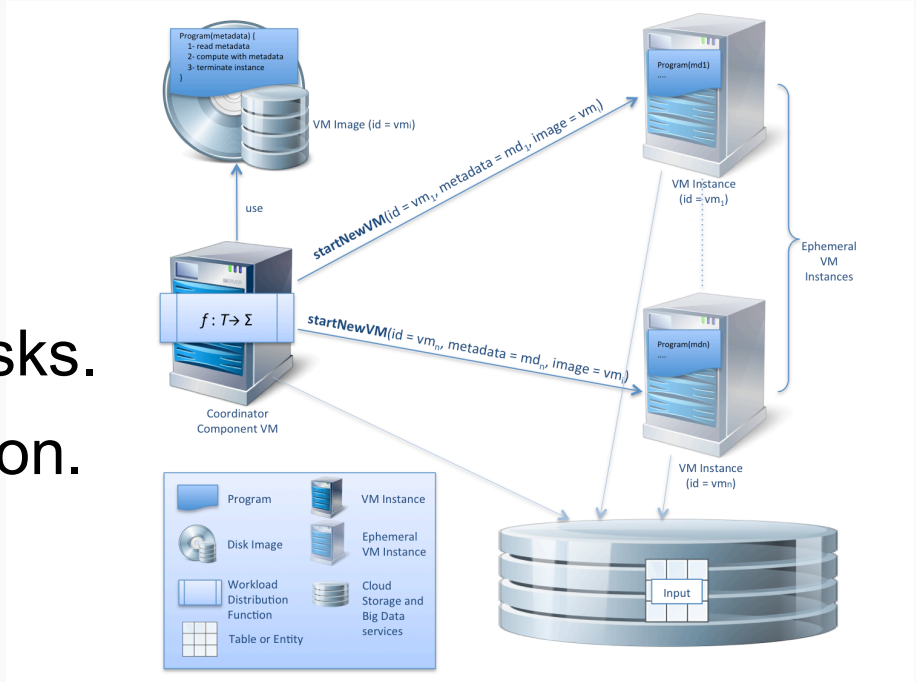
University of Reading

- Analytical big data services can analyze massive datasets in seconds (e.g. 1 terabyte in 50s).
- Can handle the analysis and storage of textual based structured and semi-structured big data.
- Data curation, transformation and normalization can be handled using an entirely parallel approach.
- Some tasks do not naturally fit the MapReduce paradigm (map/reduce, task chaining, complex logic, data streaming).
- Frameworks such as Hadoop utilizes a fixed number of computing nodes during processing.
- Cloud computing elasticity can be utilized to scale up and down VMs as needed.

Major Components

University of Reading

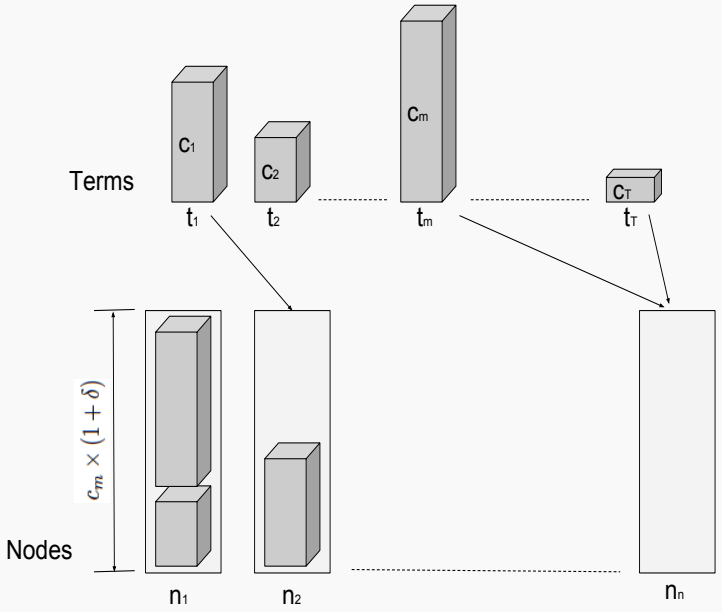
- Coordinator VM.
- Processor VMs.
- Processor VM Disk Image
- Job/Work description.
- Processing program and tasks.
- Workload Distribution function.
- Cloud storage.
- Analytical big data service.
- Program input via VM metadata.



Workload Distribution

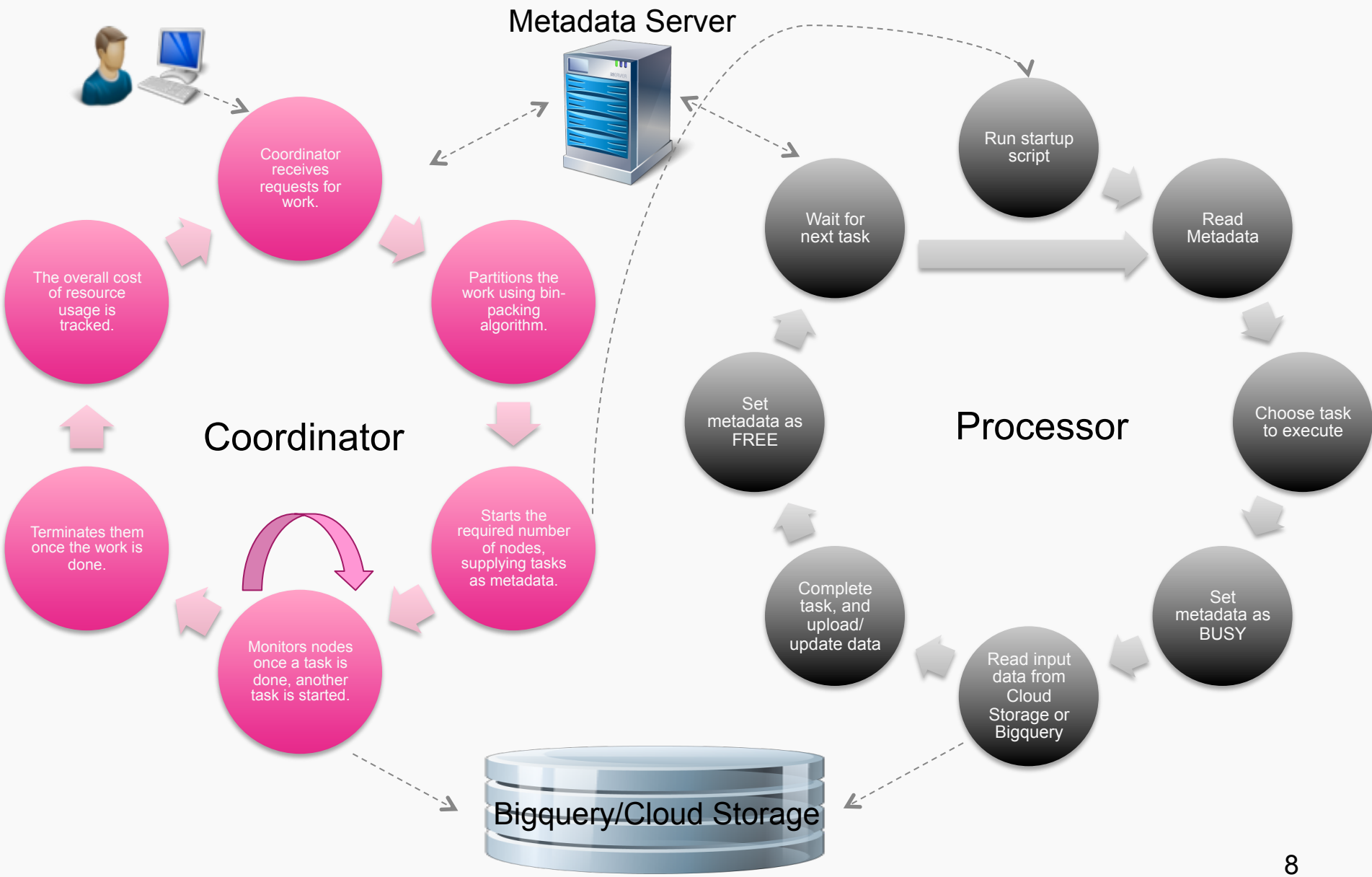
University of Reading

- Task processing is entirely parallel, so processors do not need to communicate with each other.
- Work is distributed using bin packing to ensure each processor is fairly loaded.
- Items to partition can be files to process or analytical queries to run against Bigquery.



Processing Steps

University of Reading

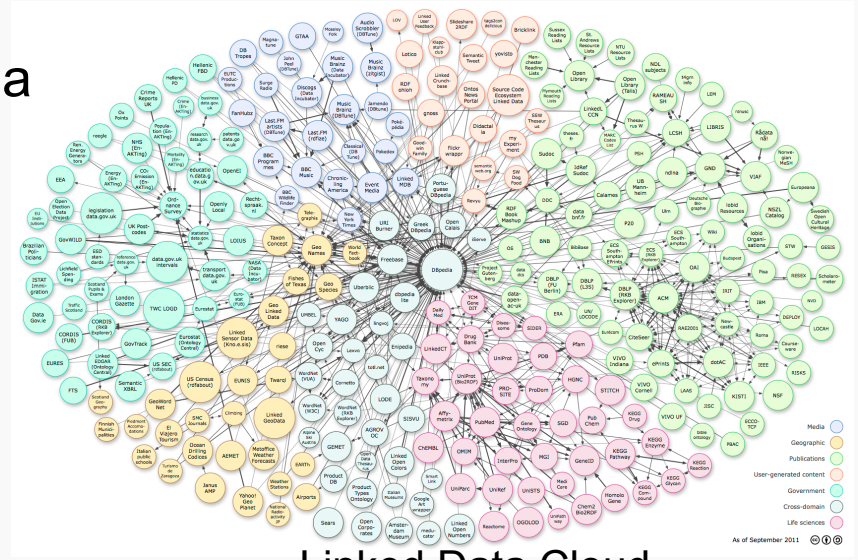


Experiment

University of Reading

- Experiment conducted on the Google Cloud Platform:
  - Compute Engine: Up to 10 processors of type n1-standard2 VMs each with 2 virtual cores, 10 GB disk and 7.5 GB of main memory.
  - Cloud Storage
- DBpedia* dataset is used:
  - Structured extract from Wikipedia
  - Contains 300 Million statements
  - Total size is 50.19 GB
  - Compressed size is 5.3GB
  - Data is in NTriple RDF format:

<http://dbpedia.org/resource/AccessibleComputing>
<http://xmlns.com/foaf/0.1/isPrimaryTopicOf>
<http://en.wikipedia.org/wiki/AccessibleComputing> .

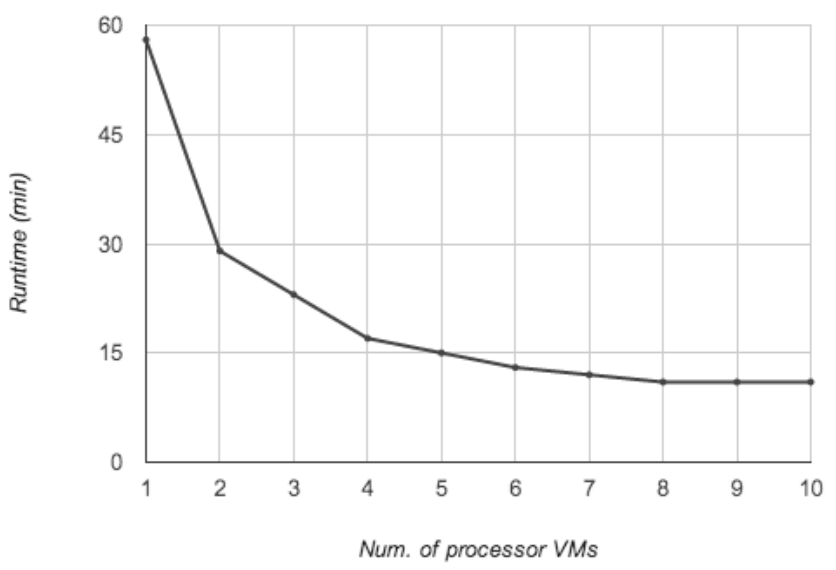


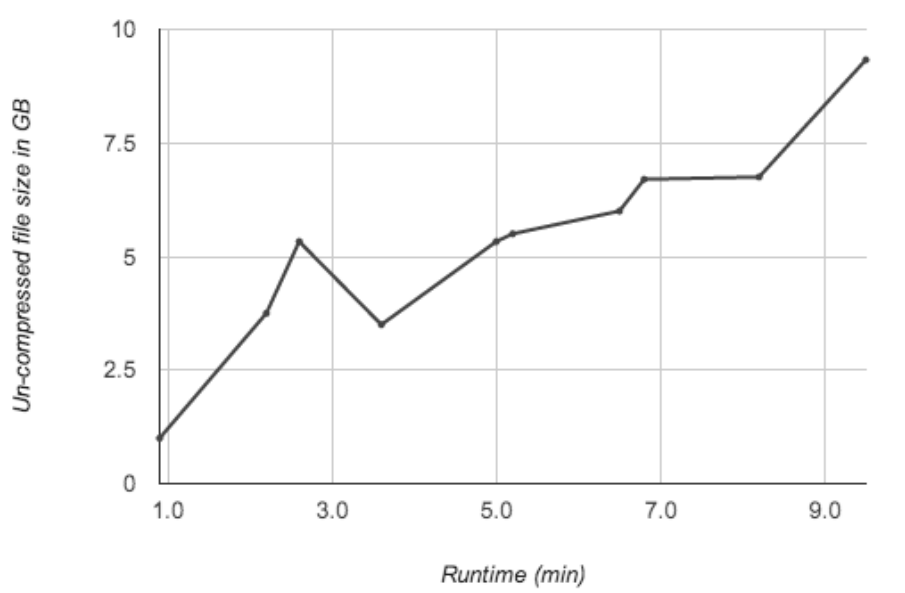
Linked Data Cloud
http://lod-cloud.net/versions/2011-09-19/lod-cloud_colored.html

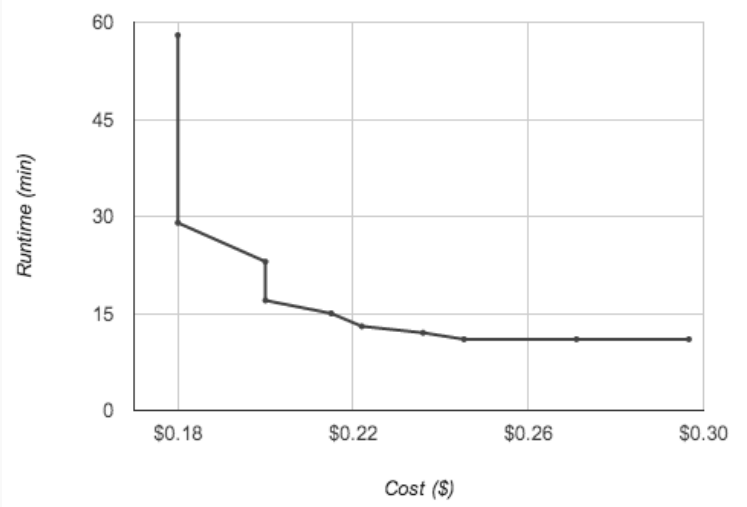
* http://wiki.dbpedia.org/Datasets

Results

University of Reading







Discussion

University of Reading

- Preprocessed 50GB of data in 11 minutes using 8 VMs.
- For our data, the processing is CPU bound (80% processing, 20% I/O).
- Processing time is proportional to the size of the data assigned to the VM.
- The overall runtime is constraint by the time required to process the largest file.
- Input files can be split further to enable equal workload allocation.
- Only 9% to 20% of the overall runtime is spent in transferring the files to and form cloud storage.

Conclusion and future work

University of Reading

- We have developed a novel cloud based framework for Big Data preprocessing.
- Our framework is lightweight, elastic and extendible.
- Makes use of cloud storage and analytical big data services to provide a complete pipeline for big data processing.
- We have extended the processing to executing analytical queries against Bigquery.
- We plan to use the framework for processing social media datasets.
- The implementation for our framework is open source and can be downloaded from http://ecarf.io

Thank you

University of Reading

- Any questions?