SANCTUM

Engineering Project: Scalable Analysis of Twitter data on a Cluster Using Mining techniques.

SANCTUM

Agenda:

General:

1. Project Description
2. System Diagram
3. Dataset and Sample Data
4. Apache Hadoop

Detailed:

5. Information Retrieval
6. Association Rule Mining
7. Interface and Visualization

General:

8. Ethics
9. Anticipated Outcomes
10. Risk Matrix
11. Project Plan

1. Project Description

Background:

Benefits of analysing social media data:

a. Gauging user sentiment for proactive planning.
b. Developing effective recommendation systems.
c. Identify influential individuals in online communities.
d. Discovering interesting associations.

Problems related to the analysis of social media data:

a. Usually complex, unstructured format.
b. Enormous amounts of data.

1. Project Description

Problem:

The need to generate query results and extract recurring patterns from a huge set of Twitter data in a timely manner.

Solution:

We aim to build a system that can utilize a cluster environment (ICTS cluster) to do a scalable analysis of the Twitter data.

1. Project Description

3-part system:

1. A Parallel Information Retrieval engine (Matt Young)

2. A Parallel Association Rule Mining algorithm (Eric Dai)

3. A web-based UI with useful visualizations (Pieter vd Walt)

1. Project Description

Preliminary Requirements:

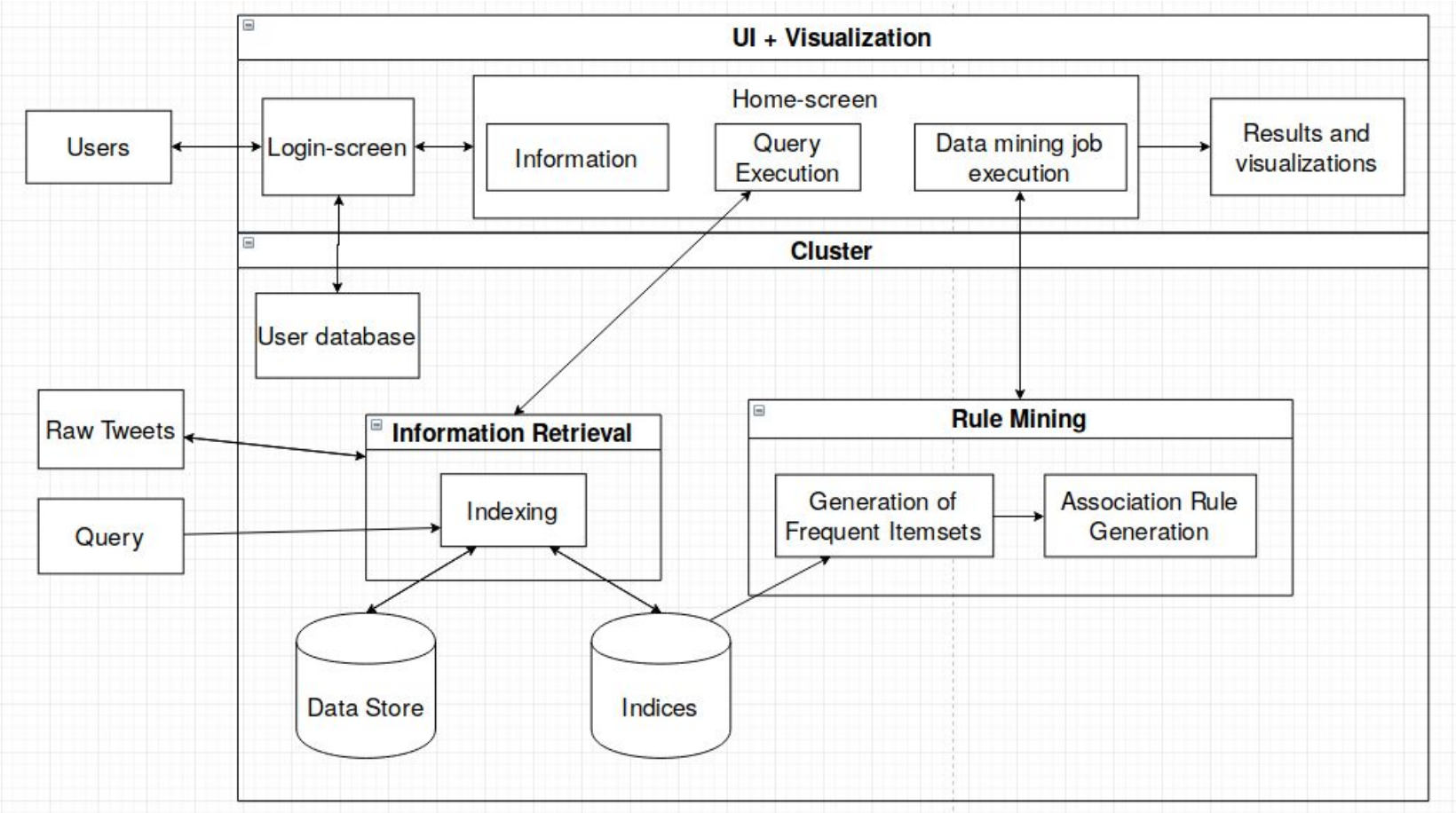
Back-end:

1. Provide pipelining capabilities for staging, preprocessing and analysis.
2. Parallel Information retrieval engine.
3. Parallel Association Rule Mining algorithm.
4. Expectations of jobs to complete in an order of minutes. Very dependant on job size.

Front-end:

1. Homepage should display hardware configurations, estimated completion time for jobs, log of previously completed jobs.
2. Provide a method for ring fencing data through the execution of queries (hashtags, usernames, phrases, individual words).
3. Display rule mining results with computed summaries, as well as useful graphs.
4. Be able to export results in XML format.

2. System Diagram



3. Dataset and Sample Data

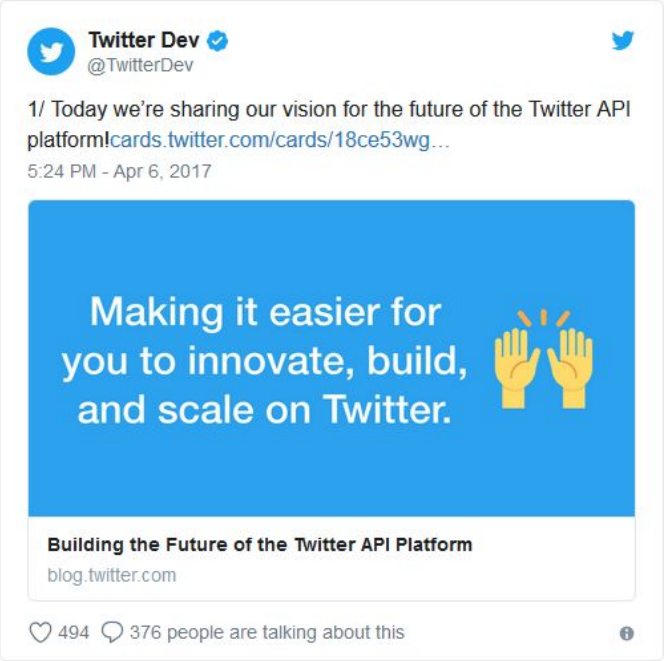
Data Source: Obtained by Prof. Hussein Suleman

Data Size: 30 Terabytes

Data Format: Twitter ‘tweets’

3. Dataset and Sample Data

Tweet in a web browser:



JSON Representation:


```
{
  "created_at": "Thu Apr 06 15:24:15 +0000 2017",
  "id_str": "850006245121695744",
  "text": "1\/ Today we\u2019re sharing our vision for the future of the Twitter API
platform!\nhttps:\/\/t.co\/XweGngmxlP",
  "user": {
    "id": 2244994945,
    "name": "Twitter Dev",
    "screen_name": "TwitterDev",
    "location": "Internet",
    "url": "https:\/\/dev.twitter.com\/",
    "description": "Your official source for Twitter Platform news, updates & events.
Need technical help? Visit https:\/\/twittercommunity.com\/ \u2328\ufe0f #TapIntoTwitter"
  },
  "place": {
  },
  "entities": {
    "hashtags": [
    ],
    "urls": [
      {
        "url": "https:\/\/t.co\/XweGngmxlP",
       "unwound": {
         "url": "https:\/\/cards.twitter.com\/cards\/18ce53wgo4h\/3xo1c",
         "title": "Building the Future of the Twitter API Platform"
       }
     }
    ],
    "user_mentions": [
    ]
  }
}
```

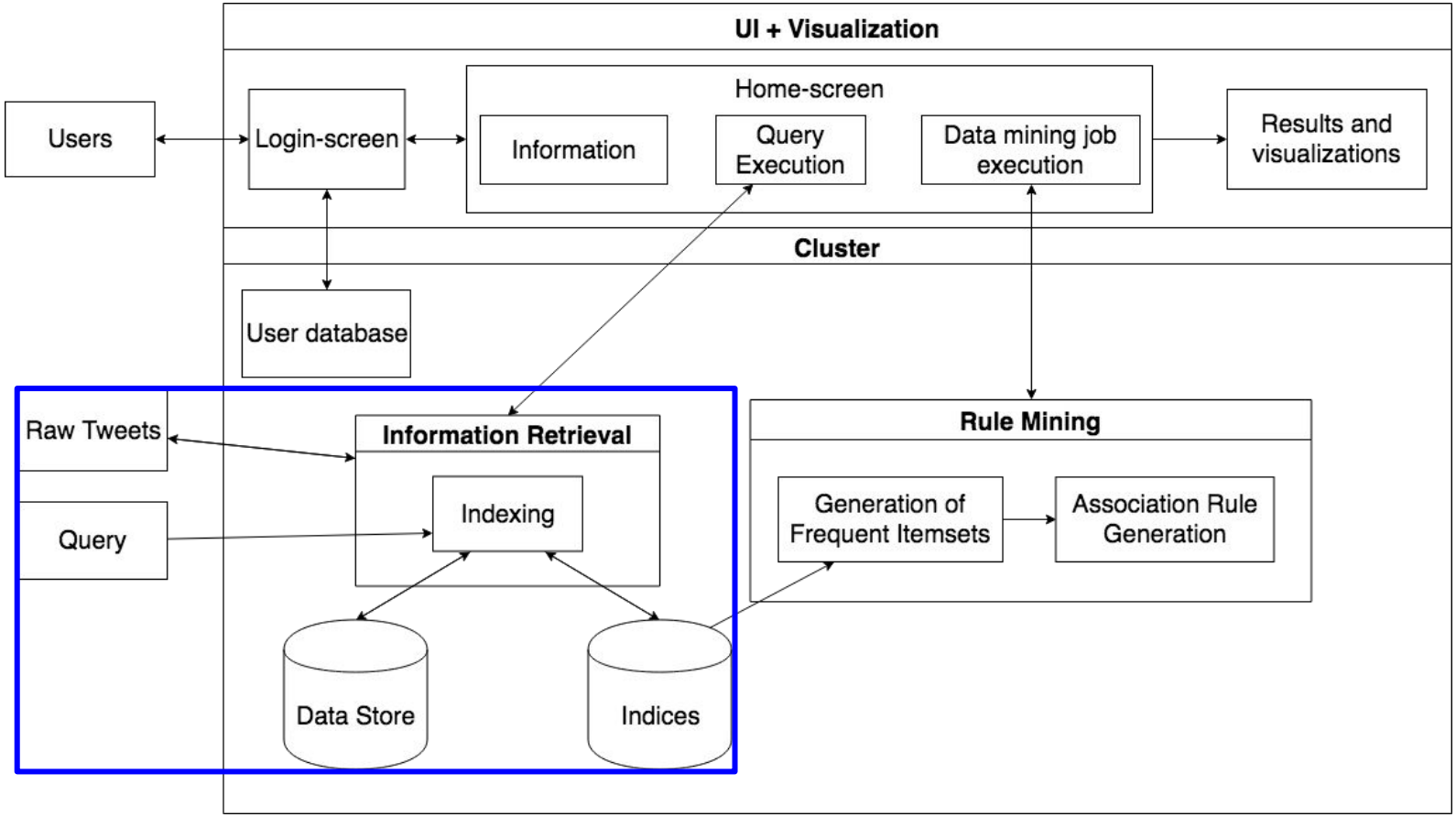
4. Apache Hadoop

- Framework of choice
- Provides a distributed programming framework for processing large amounts of data using the MapReduce programming model
- Programming in Java

4. Apache Hadoop

- Contains the following subsystems:
  - Hadoop Distributed File System (HDFS)
  - MapReduce engine (JobTracker & TaskTracker)
- Will be used for both the Information Retrieval and Association Rule Mining

5. Information Retrieval



5. Information Retrieval

Information retrieval is the finding of material of an unstructured nature from large collections that satisfies an information need

Performs three major tasks:

- Preprocessing - removing all unwanted topics from each tweet
- Querying - involves processing the user request from the front-end / user interface to fetch the desired data.
- Indexing - involves organising the data in a data structure, which can be accessed by the pattern recognition system (inverted index, ranking and scoring algorithms).

5. Information Retrieval

Evaluation:

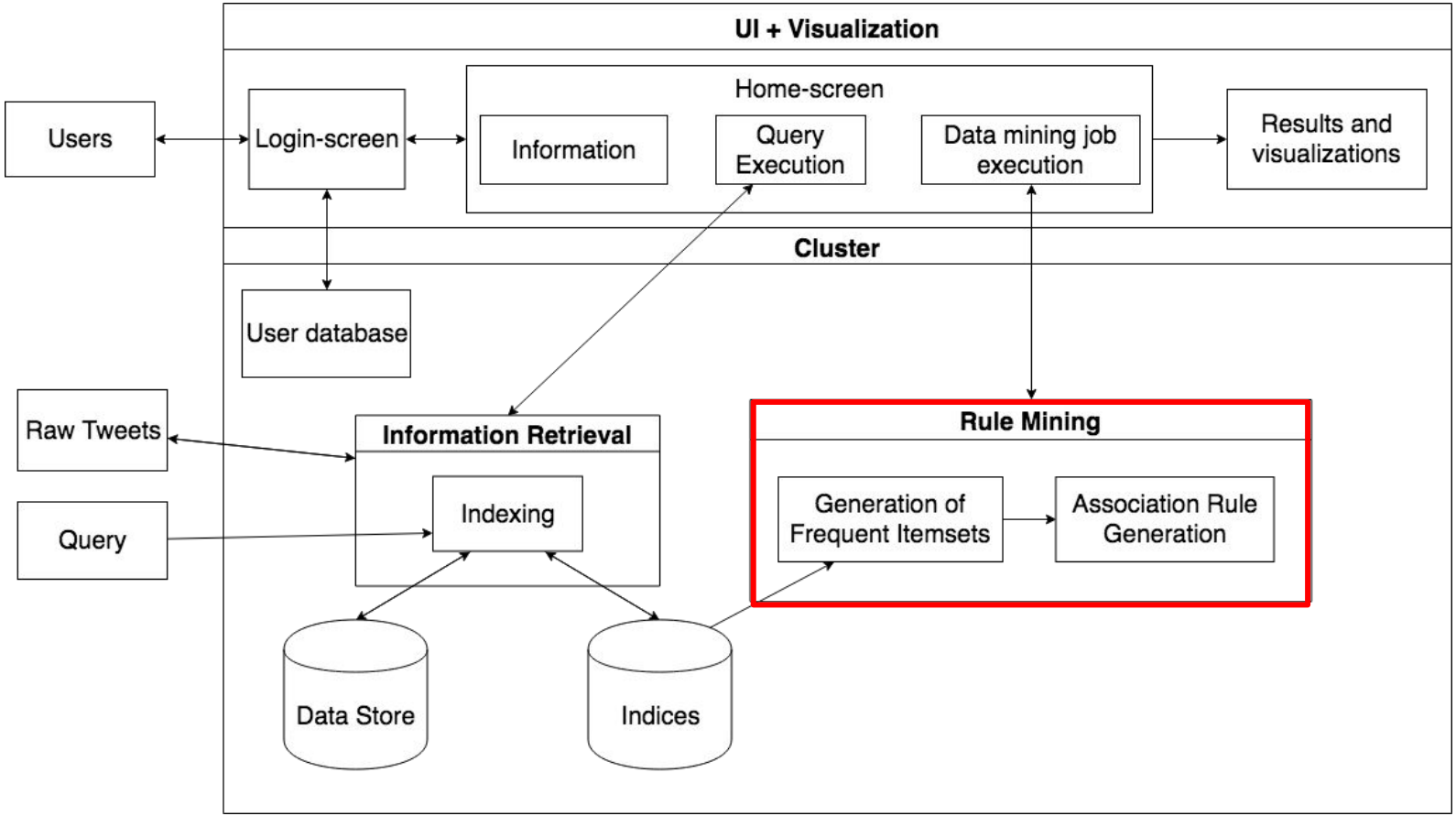
- Computation time: the overall running time to complete an information retrieval task
- Speedup: the factor at which the computation time increases as the data input size increases
- Correctness comparison: using an existing testbed of established reliable data, a comparison between this system's output and the existing output will show a degree of correctness in our system.

5. Information Retrieval

Related Work:

- Using partial replication of text databases on multiple servers [1], analysing resource allocation
- A scalable parallel information retrieval system [2] to fetch documents relevant to a specific research topic
- Scalable distributed information retrieval system for hundreds of gigabytes of data using *clients*, *connection servers*, and *Inquery servers* [3]

6. Association Rule Mining



6. Association Rule Mining

Overview:

- Generates rules between items in basket-like data
- Define:
  - Support threshold
  - Confidence threshold
- Standard Apriori Algorithm
  - Generation of frequent itemsets
  - Generation of association rules

ID	Items
1	{Bread, Milk}
2	{Bread, Diapers, Beer, Eggs}
3	{Milk, Diapers, Beer, Cola}
4	{Bread, Milk, Diapers, Beer}
5	{Bread, Milk, Diapers, Cola}
…	…

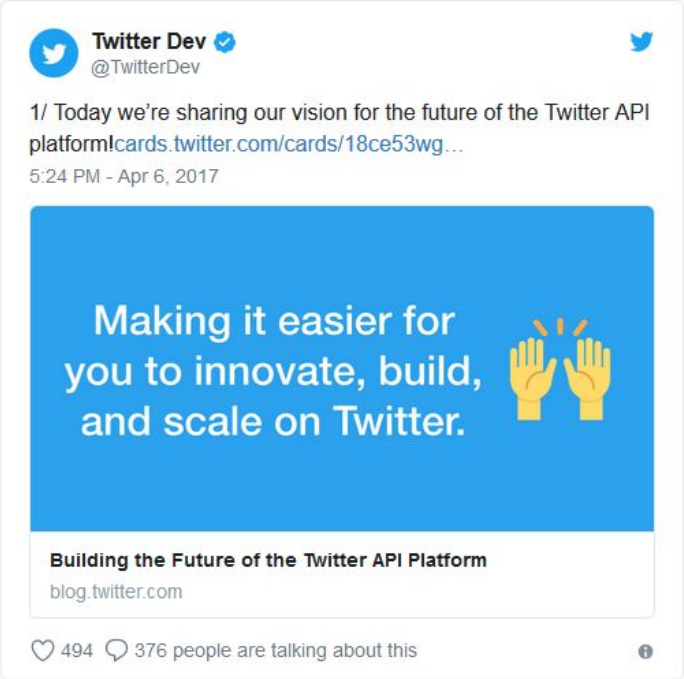
market basket transactions

{Diapers, Beer} Example of a frequent itemset

{Diapers} → {Beer} Example of an association rule

6. Association Rule Mining

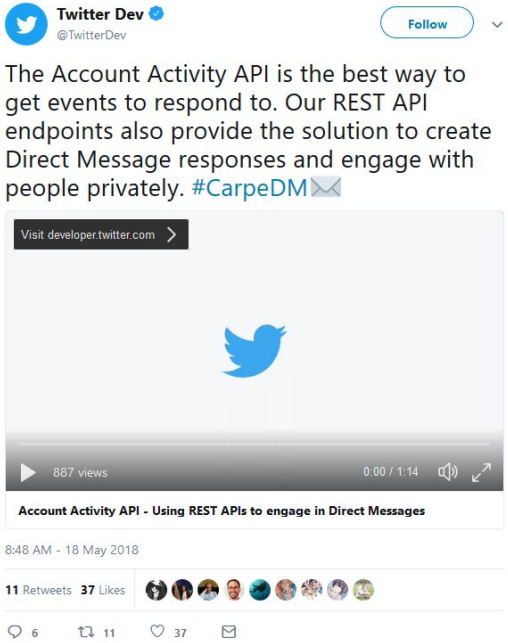
Example of basket-like tweet:




```
{"Today", "sharing", "future",
"vision", "platform", "Twitter",
"API"}
```

6. Association Rule Mining

Example of basket-like tweet:




```
{"API", "Account", "Activity", "best,
"way", "REST", "events", "respond",
"Direct", "Message", "endpoints",
"provide", "provide", "create",
"solution", "responses", "engage",
"privately", "people", "#CarpeDM"}
```

6. Association Rule Mining

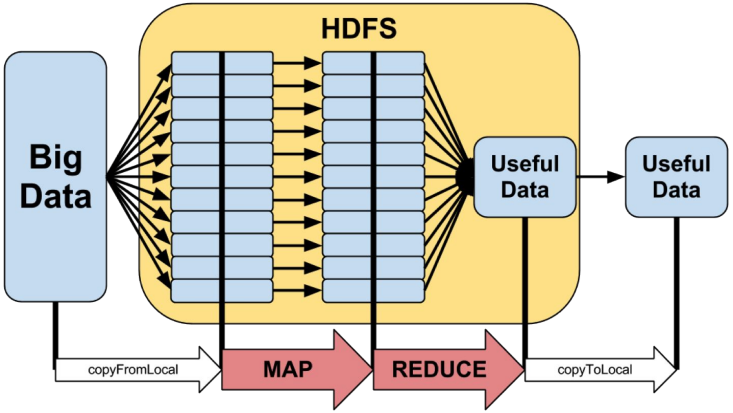
Related Work:

- MapReduce as a programming model for association rules algorithm on Hadoop [4].
  - Identified generation of candidate itemsets as slow
  - Proposed a parallel Apriori over MapReduce using Hadoop
  - Result: speedup of the algorithm increased linearly with datasets from 100,000 transactions to 1,000,000 transactions
- Efficient mining of frequent itemsets in social network data based on MapReduce framework [5].
  - Focused on rule mining social media data
  - Proposed a scalable algorithm, also parallel Apriori over MapReduce
  - Algorithm also had pruning techniques to further increase efficiency
  - Results: computation time on the order of 1000 sec. on datasets of 1,000,000 transactions

6. Association Rule Mining

Implementation:

- Implementation - Aims for scalability
  - Will be done in a cluster environment
  - Implement slower performing steps of Apriori in parallel
  - Namely, the generation of frequent itemsets onto MapReduce (using Hadoop)

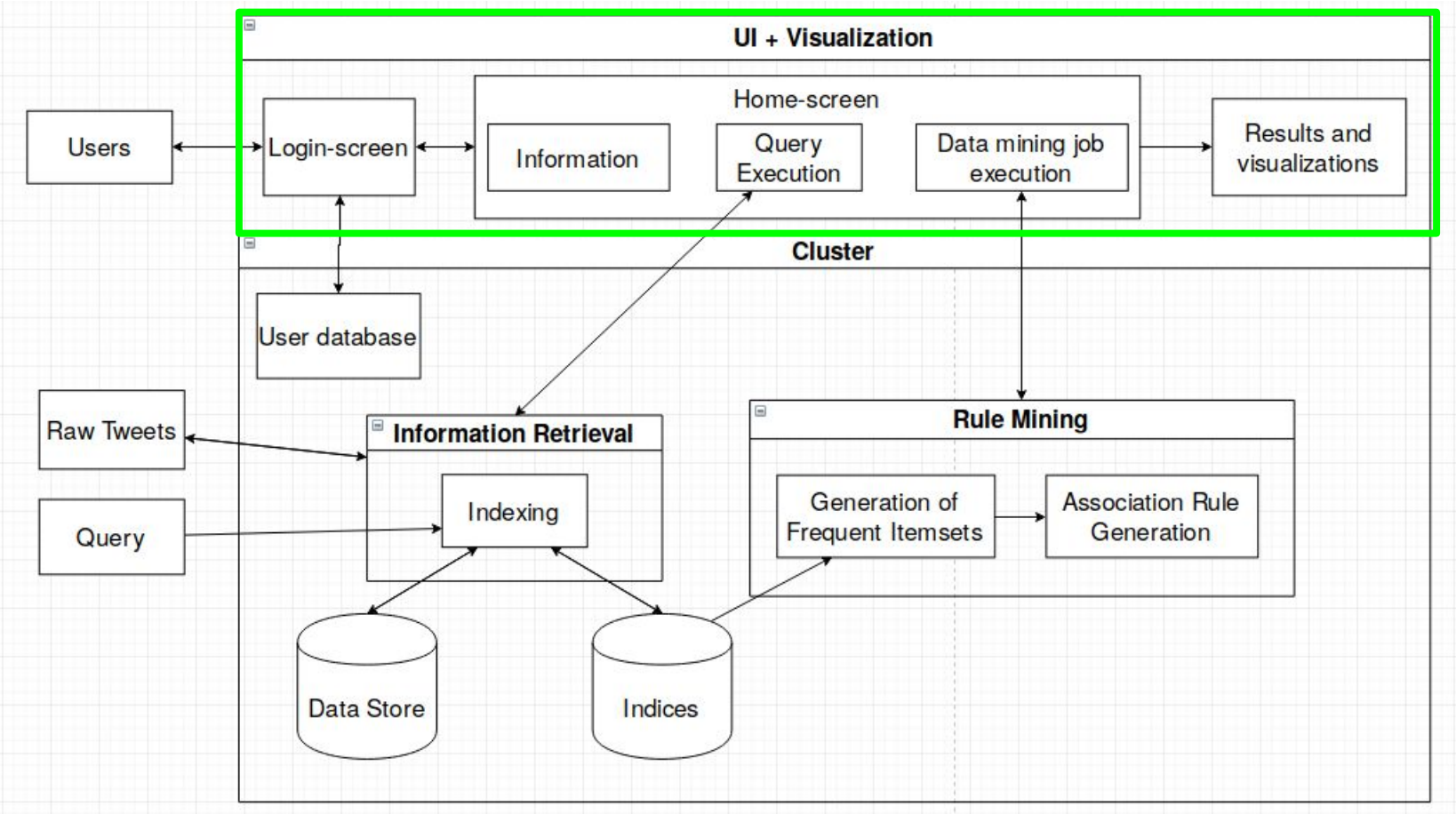


6. Association Rule Mining

Evaluation:

- Execution time as the dataset increases
- Speedup as the dataset increases
- Correctness by testing against established testbeds
    - IBM generated synthetic dataset

7. Interface and Visualizations



7. Interface and Visualizations

Framework:

ExpressJS

Free and open-source web application framework built for Node.js platform.

Chosen because of simplicity, flexibility and scalability.

Can integrate with useful libraries that will help with the visualizations (Plotly, FusionCharts, AmCharts D3.js)

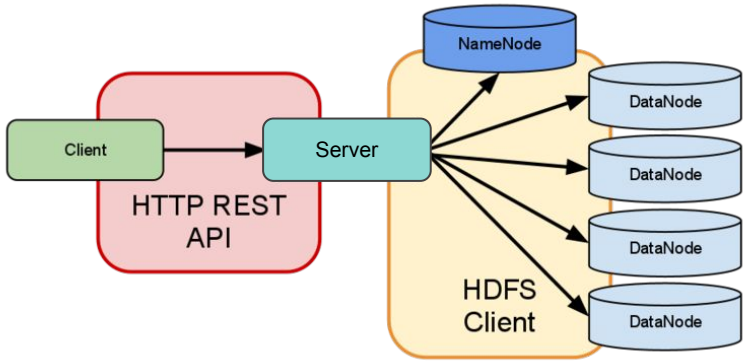
7. Interface and Visualizations

APIs:

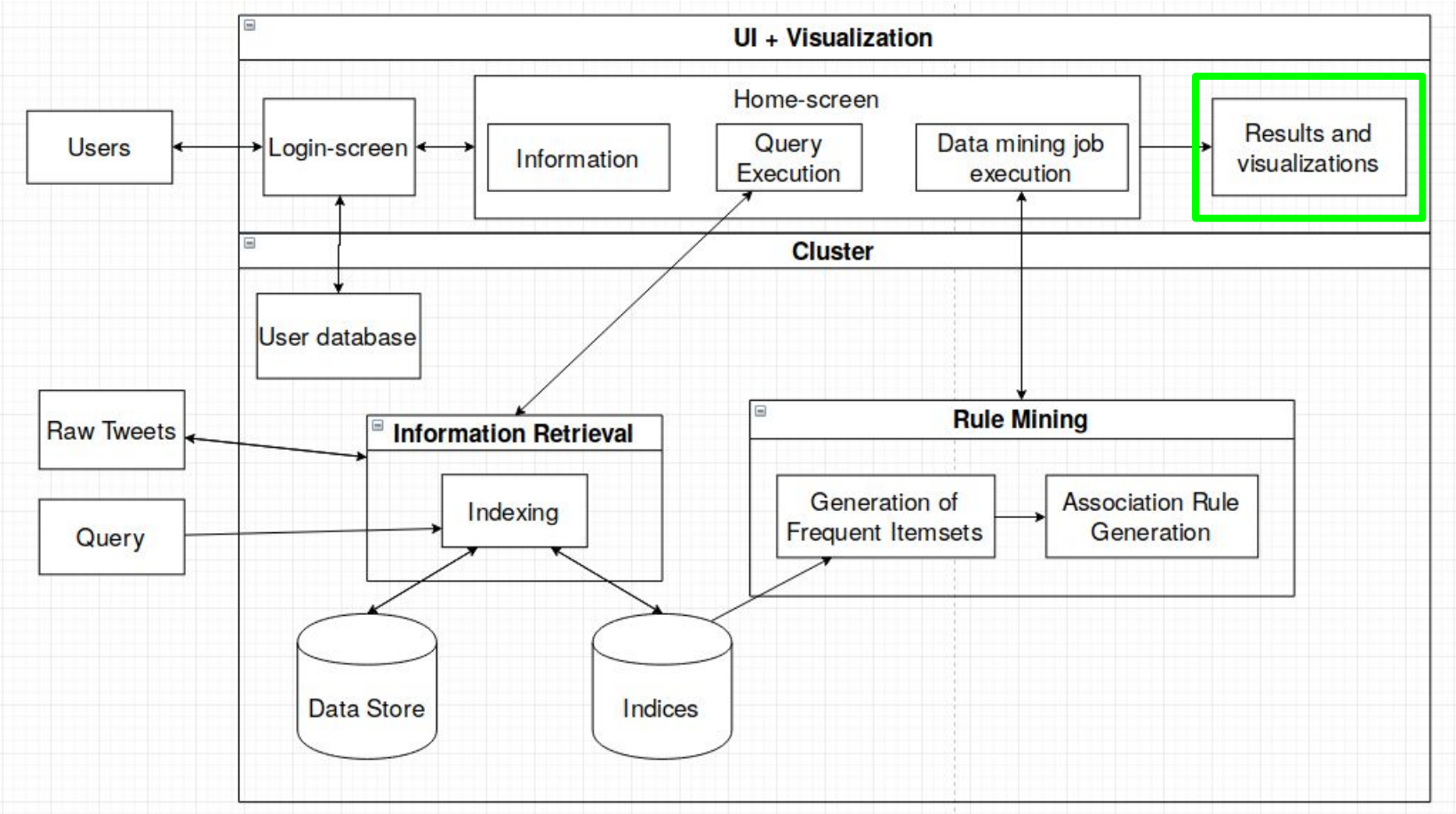
Hadoop YARN Web Service REST APIs will be used to interact with the cluster.

YARN: Resource manager and job scheduler in the Hadoop system.

REST APIs: Set of URIs that grant access to applications on the nodes on the cluster.



7. Interface and Visualizations



7. Interface and Visualizations

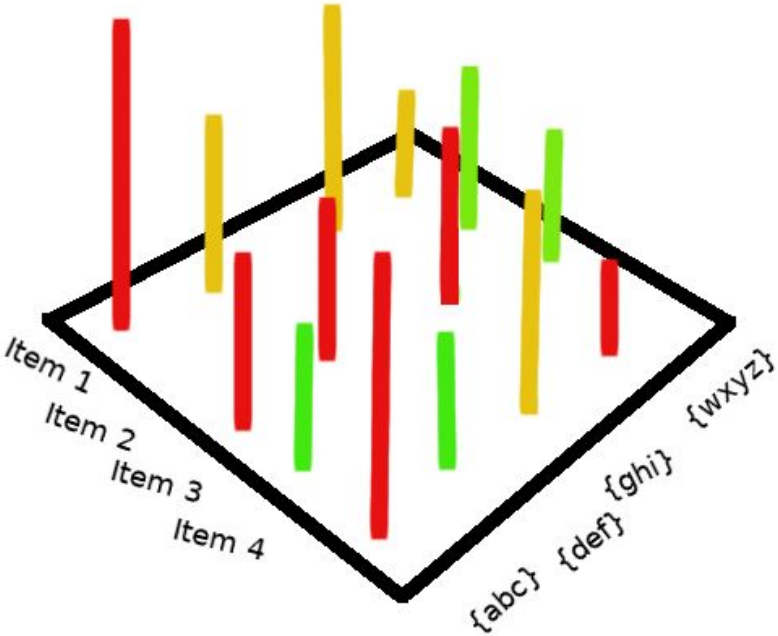
Related Work:

- Visualizing association rules in hierarchical groups [6].
  - Explored various visualization options. The most interesting one being the “grouped matrix”.
- Interactive visual exploration of association rules with rule-focusing methodology [7].
  - Introduced a very interactive 3D visualization.
- Visual Twitter analytics: Exploring fan and organizer sentiment during Le Tour de France [8].
  - Visualized the changing sentiment towards an event by analysing Twitter data.

7. Interface and Visualizations

Association rule visualizations:

1. 3D bar graph:



Confidence level

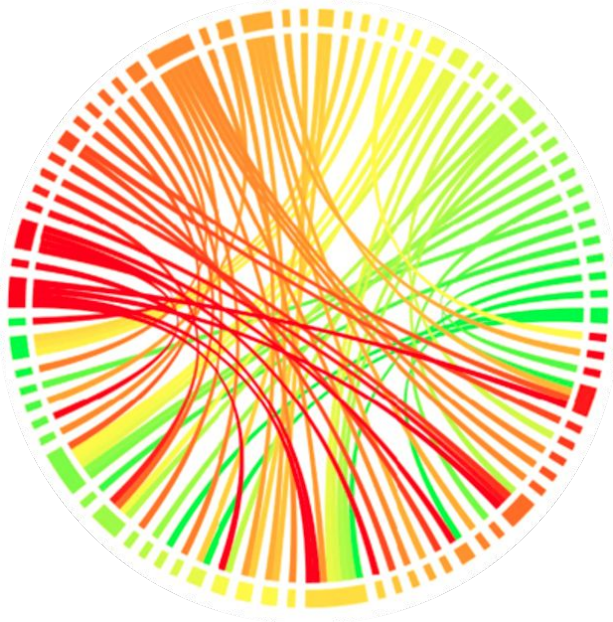
0% 100%

Height = support level

7. Interface and Visualizations

Association rule visualizations:

2. Circos plot:



Confidence level

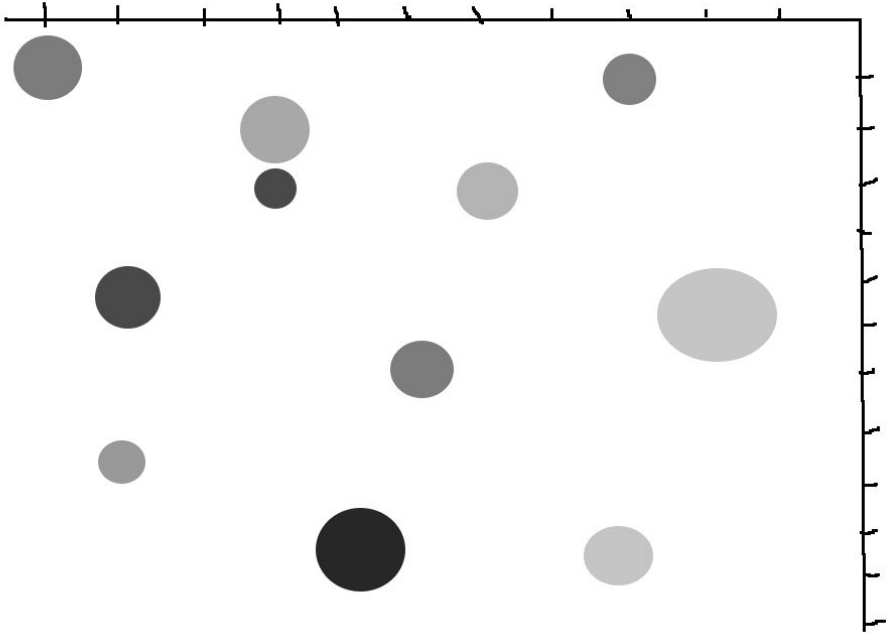
0% 100%

Thickness = support level

7. Interface and Visualizations

Association rule visualizations:

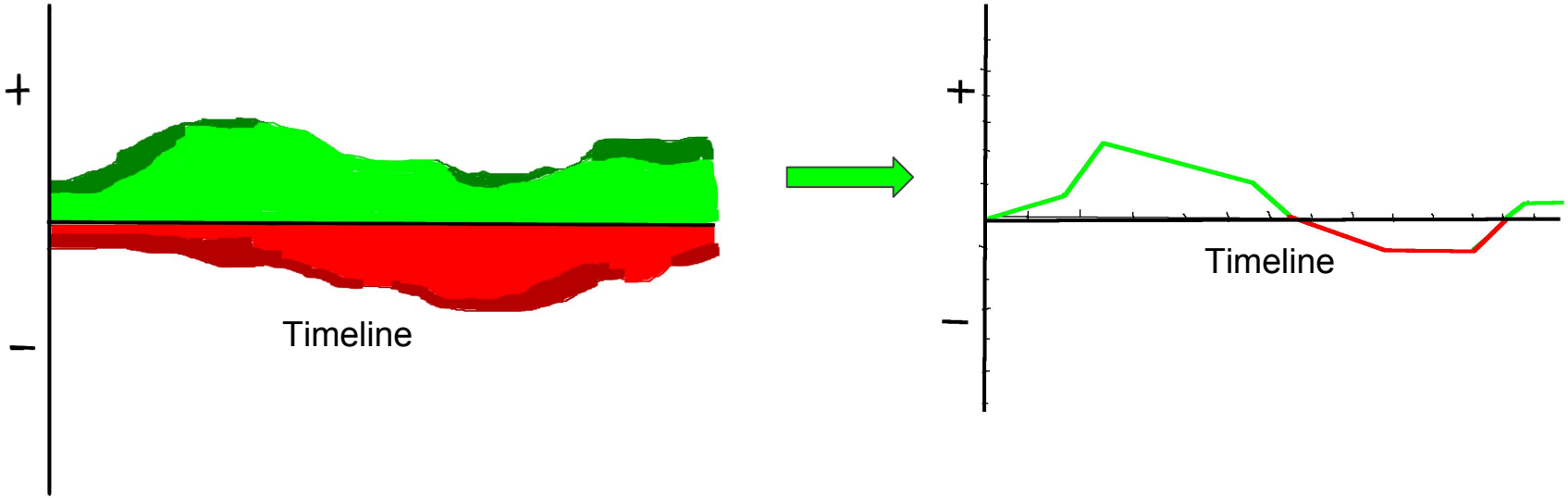
3. Grouped Matrix [6]



Circle size = support level

7. Interface and Visualizations

Sentiment analysis visualization (possibly): [8]



7. Interface and Visualizations

Evaluation:

Standard HCI evaluation techniques will be used

- Heuristic evaluation (Nielsen and Molich)
- Evaluation through user participation
- Observational methods

8. Ethics

- Social link disclosure
  - The ability to identify sensitive or private relationships that were meant to stay hidden
- Affiliation link disclosure
  - The ability to identify whether somebody is affiliated with a particular group

9. Anticipated Outcomes

- System outcomes
  - Information Retrieval module
  - Association Rule Mining module
  - Web Front-End Tool
- Key Success Factors
  - Have the algorithms been implemented successfully?
  - Does our system work efficiently and scale with large datasets in a distributed computing environment?
  - Does the system provide scalable and useful visualizations of data mining results?

10. Risk Matrix

Risk	Impact	Probability	Mitigation
Unable to meet project deadlines	High	Low	Start early, and stick to intermediate deadlines
Twitter data given is too noisy	Medium	Medium	Extra steps will have to be implemented to pre-process and clean the data
Unable to meet system requirements	High	Low	Consult supervisor on problematic areas in early iterations to discuss issues and challenges faced
UCT's HPC cluster is down for the duration of the project	Medium	Low	Apply for another cloud computing service to evaluate the system with

11. Project Plan

12. References

[1] Z. Lu, K. S. McKinley. *Searching a Terabyte of Text Using Partial Replication*. Department of Computer Science, University of Massachusetts, Amherst, MA

[2] Peter Bailey, David Hawking. 2007. *A Parallel Architecture for Query Processing Over A Terabyte of Text.* The Australian National University, Canberra, Australia

[3] Brendon Cahoon, Kathryn S. McKinley, Zhihong Lu. 2000. *Evaluating the Performance of Distributed Architectures for Information Retrieval Using a Variety of Workloads*. University of Massachusetts, Massachusetts, USA

[4] Yang, X.Y., Liu, Z. and Fu, Y., 2010, June. MapReduce as a programming model for association rules algorithm on Hadoop. In *Information Sciences and Interaction Sciences (ICIS), 2010 3rd International Conference on* (pp. 99-102). IEEE.

[5] Farzanyar, Z. and Cercone, N., 2013, August. Efficient mining of frequent itemsets in social network data based on MapReduce framework. In *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining* (pp. 1183-1188). ACM.

[6] Michael Hahsler and Radosław Karpienko. 2016. Visualizing association rules in hierarchical groups. Journal of Business Economics 87, 3 (July 2016), 317–335.

[7] Julien Blanchard, Fabrice Guillet, Henri Briand. Interactive visual exploration of association rules with rule-focusing methodology. Knowledge and Information Systems (KAIS), Springer, 2007, 13 (1), pp.43-75. <10.1007/s10115-006-0046-2>.<hal-00420938>

[8] Larena Hoeber, Laura Wood, Ryan Snelgrove, Isabella Hugel, and Dayne Wagner. 2013. Visual Twitter analytics: Exploring fan and organizer sentiment during Le Tour de France. In Proceedings of the IEEE VIS Workshop on Sports Data Visualization. IEEE.

Questions?