Apache Spark: A Literature Review

Presenter: Aaron Sarson
Outline

- Introduction to Spark
- Problem to be addressed
- Proposed Approach
  - Research Questions
- Contributions
- Results
  - RQ1, RQ2, RQ3
- Conclusion & Future Work
Introduction to Apache Spark

- Apache Spark is an open source big data processing framework.
- Spark has become increasingly popular due to:
  - Avoids I/O bottlenecks
  - In-memory computations
- Uses lineage-based recovery which is more efficient recovery method employed in Hadoop.
- Apache Spark has a streaming component useful for providing input for real time analytics.
Problem

Lack of systematic survey that focuses on Apache Spark.
  - What is state of the art work being done with Spark?

Perform the gap analysis based on current literature.
  - Discover limitations within current literature about Apache Spark
  - Provide direction to researchers where the unexplored areas are

i. No analysis of the domains where Apache Spark is being utilized
ii. Lack of examination into the types/categories of papers
iii. No exploration of the tools/components that have been used with Spark
Proposed Approach

Number of papers to collect: 20
  - These papers will should represent a \textit{snapshot} of the research area

Databases Searched: ACM Digital Library, IEEE Library, Google Scholar

Query Strings: \{“spark”, “apache spark”, “spark-based”\} x
  \{“”, “health”, “networking”, “BI”, “system”, “prediction”\}

Inclusion/exclusion criteria:
- ✔ Journal articles, conference papers/proceedings
- ✔ Publication Year: [2014, 2016]
- ✗ Books, articles which only briefly mention spark
Proposed Approach- Research Questions

The questions this literature review is considering:

RQ1. What domains does the literature about Spark encompass?
RQ2. What categories of papers are in literature?
RQ3. What tools are used in the literature with Apache Spark?
Contributions

- My contributions to this research problem:
  - Systematic literature review on the current literature about Apache Spark
  - Literature review will evaluate the state of the art research as well as present a gap analysis
## Results – RQ1

<table>
<thead>
<tr>
<th>Domain</th>
<th>Count</th>
<th>1.</th>
<th>2.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Health</td>
<td>2</td>
<td>“A Spark-based workflow for probabilistic record linkage of health care data”</td>
<td>“SparkGIS: Efficient Comparison and Evaluation of Algorithm Results in Tissue Image Analysis Studies”</td>
</tr>
<tr>
<td>E-Commerce</td>
<td>1</td>
<td>“A Reference Architecture and Road map for Enabling E-commerce on Apache Spark”</td>
<td></td>
</tr>
<tr>
<td>Media</td>
<td>2</td>
<td>“A Spark-based Political Event Coding”</td>
<td>“Real-time News Recommendations using Apache Spark”</td>
</tr>
<tr>
<td>Network Monitoring</td>
<td>1</td>
<td>“A Cloud Computing Based Network Monitoring and Threat Detection System for Critical Infrastructures”</td>
<td></td>
</tr>
<tr>
<td>Other</td>
<td>1</td>
<td>“Apache Spark: A Unified Engine for Big Data Processing”</td>
<td></td>
</tr>
</tbody>
</table>
## Results - RQ2

<table>
<thead>
<tr>
<th>Category</th>
<th>Count</th>
<th>1.</th>
<th>2.</th>
<th>3.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Records Management</td>
<td>1</td>
<td>“A Spark-based workflow for probabilistic record linkage of health care data”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Natural Language Processing</td>
<td>1</td>
<td>“A Spark-based Political Event Coding”</td>
<td></td>
<td></td>
</tr>
<tr>
<td>NoSQL document database</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
<td>--------------------------</td>
</tr>
<tr>
<td>MongoDB</td>
<td>CouchDB</td>
<td>Terrastore</td>
<td>MongoDB</td>
<td>CouchDB</td>
</tr>
<tr>
<td>NoSQL key-value database</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>S3</td>
<td>Cassandra</td>
<td>S3</td>
<td>Cassandra</td>
<td>S3</td>
</tr>
<tr>
<td>NoSQL columnar database</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>HBASE</td>
<td>Riak</td>
<td>Cassandra</td>
<td>HBASE</td>
<td>Riak</td>
</tr>
<tr>
<td>Hadoop (HDFS)</td>
<td>Apache Spark</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Core</td>
<td>SQL</td>
<td>Streaming</td>
<td>MLib</td>
<td>GraphX</td>
</tr>
<tr>
<td>SQL</td>
<td>Streaming</td>
<td>MLib</td>
<td>GraphX</td>
<td>CRM suite</td>
</tr>
<tr>
<td>Streaming</td>
<td>MLib</td>
<td>GraphX</td>
<td>CRM suite</td>
<td>Standford CoreNLP</td>
</tr>
<tr>
<td>MLib</td>
<td>GraphX</td>
<td>CRM suite</td>
<td>Standford CoreNLP</td>
<td>PERTRARCH</td>
</tr>
<tr>
<td>GraphX</td>
<td>CRM suite</td>
<td>Standford CoreNLP</td>
<td>PERTRARCH</td>
<td>MySQL database</td>
</tr>
</tbody>
</table>

Results – RQ3
Conclusion & Future Work

Preliminary Results:

- **RQ1** \(\rightarrow\) Domains where Apache Spark is used are concentrated in Health and Media
- **RQ3** \(\rightarrow\) A trend is emerging that the Apache Spark’s streaming component is a heavily utilized component
- **RQ3** \(\rightarrow\) There are not a great deal of papers using Apache Spark’s GraphX component.

Future Work:

- **Current state**: 7 papers. **Goal**: 20 papers.
- Graph and analyze the results of RQ3 in more detail
- Try and find papers in other domains such as finance/economics relating to Spark
References


References


