Applying Stratosphere for Big Data Analytics

Marcus Leich*[1], Jochen Adamek*[2], Moritz Schubotz*[3],
Arvid Heise†[4], Astrid Rheinländer‡[5], Volker Markl*[6]
*Technische Universität Berlin, Germany †Hasso-Plattner Institute Potsdam, Germany
‡Humboldt-Universität zu Berlin, Germany
[1,2,3,6]{marcus.leich,j.adamek,schubotz,volker.markl}@tu-berlin.de
[4]arvid.heise@hpi.uni-potsdam.de, [5]rheinlae@informatik.hu-berlin.de

Abstract: Analyzing big data sets as they occur in modern business and science applications requires query languages that allow for the specification of complex data processing tasks. Moreover, these ideally declarative query specifications have to be optimized, parallelized and scheduled for processing on massively parallel data processing platforms. This paper demonstrates the application of Stratosphere to different kinds of Big Data Analytics tasks. Using examples from different application domains, we show how to formulate analytical tasks as Meteor queries and execute them with Stratosphere. These examples include data cleansing and information extraction tasks, and a correlation analysis of microblogging and stock trade volume data that we describe in detail in this paper.

1 Introduction

Analytics in numerous economic, political, and scientific sectors are focussing more and more on gaining new knowledge from web scale data. Tasks like social media analysis or market monitoring require complex processing chains that need to handle structured and unstructured data. While systems such as Hadoop are capable of handling such workloads, they are comparatively difficult to develop for. Script languages like Pig Latin and JAQL have been proposed to ease development for Hadoop and facilitate ad hoc queries. This paper demonstrates the application of Stratosphere to complex analytics tasks. We present queries for three different use cases ranging from business analytics to information extraction and information integration using Stratosphere's query language Meteor. The paper is structured in the following way: Section 2 provides an overview of Stratospheres key components. Section 3 showcases how Stratosphere can be used correlate tweets with stock trade volume data. The description includes the query, it's execution, and a customized user interface. Additionally, we outline two further tasks that involve biomedical information extraction and integration of open government data. Section 4 concludes this paper.

2 Overview of Stratosphere

This section briefly describes the key components of Stratosphere relevant for the demonstration. For a more detailed description, please refer to [WK09, BEH$^+$10, HRL$^+$12].

Stratosphere[BEH$^+$10] is a massively parallel data processing system. It consists of a declarative query language **Meteor**[HRL$^+$12], the **Pact** programming model, and **Nephele** [WK09], the execution engine. Users express their queries using the Meteor language. Here, high level operators, such as filter, are applied to (semi-)structured data sets. Meteor operators consist of Pact programs, directed acyclic graphs of second-order functions. The Pact programming model is a generalization of the MapReduce concept. In addition to the conventional *map* and *reduce*, Pact provides three supplemental, second order functions which allow for efficient implementation of cross products, equi-joins, and groupings from two sources. Pact programs are optimized and compiled into data flow graphs, which are processed in parallel by the Nephele execution engine. Therefore Stratosphere is capable of transforming complex user queries into optimized parallel execution graphs.

3 Demonstrations

3.1 Correlation of Tweets and Stock Trade Volume

Objective: The program is inspired by the work of Ruiz et al. [RHC$^+$12] and computes the correlation of microblog posts (tweets) and stock trade volume. While the socioeconomic implications of this relationship are certainly interesting, this paper covers only the implementation of such an analysis.

The Meteor script: The first part of the program (Figure 1 top) specifies the sources for the tweet and stock volume data (line 3-4). Line 6 filters the relevant tweets. The following two blocks of code group the filtered tweets and trade volume data by week and aggregate the values in each group. The last portion of the script joins the tweet counts with the stock volume data, computes the correlation, and stores the result.

UI and Execution: Since the Meteor program is executed on a cluster, a web interface is provided to trigger the computation from remote machines. The UI (Figure 1) provides an input field for the query. After submission of the Meteor program, the server checks the syntax, builds the Sopremo operator graph, and compiles it into a Pact program. Stratosphere optimizes this Pact program, and executes it on Nephele. Figure 1 shows the optimized Pact program (bottom left) and the Nephele graph during execution (bottom right). After execution, result files, up to a certain size, can be inspected in the web interface. Suitable data types, like time series, can be visualized directly in the browser.

3.2 Further Applications

Finding relationships between drugs and genes is a fundamental task in pharmacogenetics, where differing drug responses due to genetic variations are studied. We present a query which extracts relationships between genes and drugs from the biomedical literature using text mining methods. First, the query analyzes the syntactical structure of the given texts and identifies occurrences of gene and drug names. Finally, a relation extraction component inspects all gene/drug name pairs, occurring in the same sentence, to detect relationships between genes and drugs.

We additionally demonstrate the integration of Open Government Data and other freely available data sets with Stratosphere. Specifically, we combine the publicly available spending from the US government to legal entities with information from Freebase about companies and their employees as well as persons and their relationships to find potential cases of nepotisms and other suspicious money flows. The results may be used by data journalists to start in-depth investigations on the involved entities.

4 Conclusion

In this paper we have demonstrated the application of Stratosphere to a correlation analysis of microblogging and stock trade volume data using Meteor.

Acknowledgements: This work is funded by the German Research Foundation under grant "FOR 1036: Stratosphere – Information Management on the Cloud" and by the European Commission under FP7 Project No. 296448 – "Data Supply Chains for Pools, Services and Analytics in Economics and Finance."

References

[BEH$^+$10] Dominic Battré, Stephan Ewen, Fabian Hueske, Odej Kao, Volker Markl, and Daniel Warneke. Nephele/PACTs: A Programming Model and Execution Framework for Web-Scale Analytical Processing. In *Proceedings of the 1st ACM symposium on Cloud computing*, SoCC '10, pages 119–130, New York, NY, USA, 2010. ACM.

[HRL$^+$12] Arvid Heise, Astrid Rheinländer, Marcus Leich, Ulf Leser, and Felix Naumann. Meteor/Sopremo: An Extensible Query Language and Operator Model. In *Proceedings of the International Workshop on End-to-end Management of Big Data (BigData) in conjunction with VLDB 2012*, Istanbul, Turkey, 0 2012.

[RHC$^+$12] Eduardo J Ruiz, Vagelis Hristidis, Carlos Castillo, Aristides Gionis, and Alejandro Jaimes. Correlating financial time series with micro-blogging activity. In *WSDM '12: Proceedings of the fifth ACM international conference on Web search and data mining*. ACM Request Permissions, February 2012.

[WK09] Daniel Warneke and Odej Kao. Nephele: efficient parallel data processing in the cloud. In *Proceedings of the 2nd Workshop on Many-Task Computing on Grids and Supercomputers*, MTAGS '09, pages 8:1–8:10, New York, NY, USA, 2009. ACM.

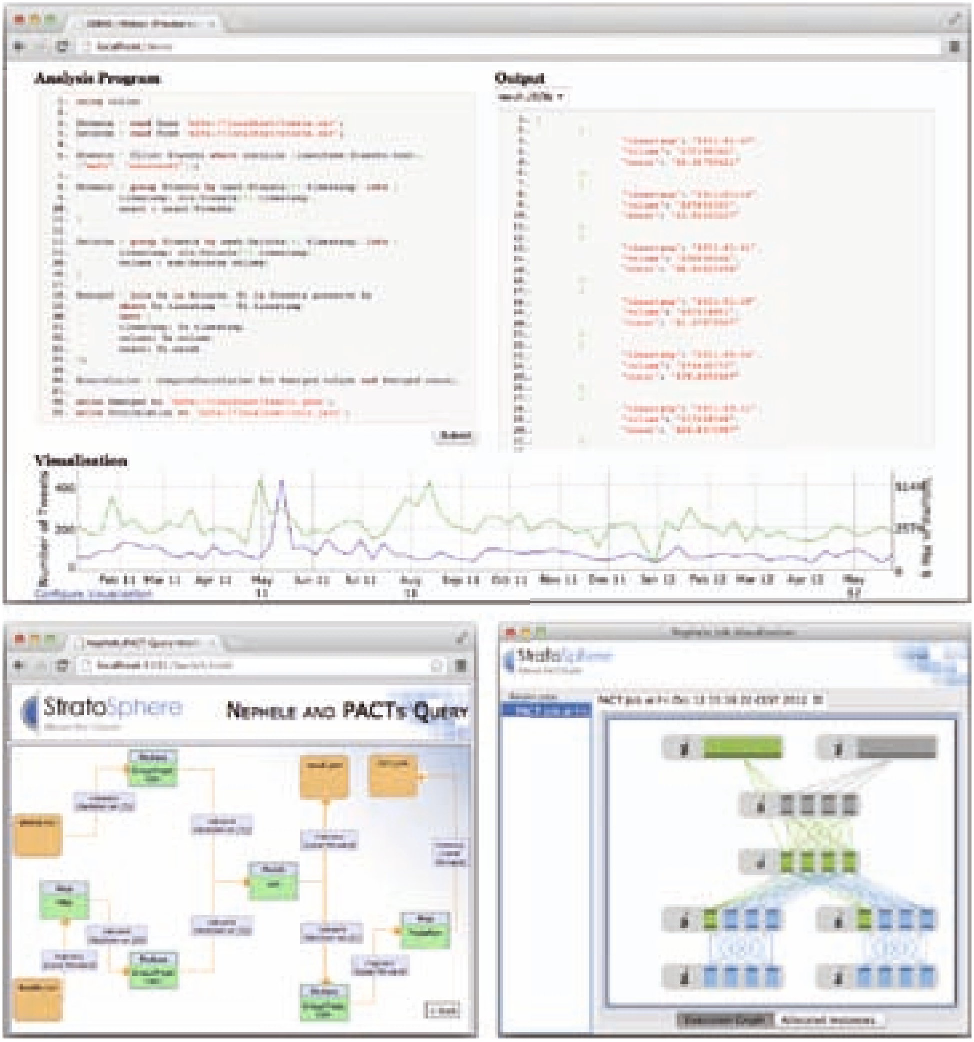

Figure 1: **Top:** Meteor's Query Submission Interface. **Bottom left:** Optimized PACT plan. **Bottom right:** Nephele execution graph.