Wochenbericht 3

Analyse und Nachverarbeitung großer wissenschaftlicher Datenmengen mit Big-Data-Tools

Arne Struck

Universität Hamburg, Fachschaft Informatik, Abschlussarbeitenseminar

3. Juni 2015

Was ist passiert?

Plan:

- Thema wählen
- Exposé nochmal anfangen
- mit WR Ausarbeitung des Themas
- Literatursuche
- Einarbeitung in das Thema

Was ist passiert?

Plan:

- Thema wählen
- Exposé nochmal anfangen
- mit WR Ausarbeitung des Themas
- Literatursuche
- Einarbeitung in das Thema

umgesetzt:

- Thema gewählt (sort of)
- Einarbeitung in Hadoop
  - Aufsetzen eines 1 Node Clusters
  - Einrichtung von Pig
  - Einrichtung von Hive
- Ein wenig damit gearbeitet
- Exposé angefangen
- Literatursuche (leider nicht all zu erfolgreich)
- SciDB angesehen

Apache Hadoop

Freies Java-Framework zur Berechnung von parallelisierbaren Problemen auf verteilten Systemen. Die Verteilung erfolgt halbautomatisiert.

Apache Hadoop

Freies Java-Framework zur Berechnung von parallelisierbaren Problemen auf verteilten Systemen. Die Verteilung erfolgt halbautomatisiert.

Hauptkomponenten:

- HDFS (Hadoop distributed file system)
- MapReduce Implementation

Apache Hadoop

Freies Java-Framework zur Berechnung von parallelisierbaren Problemen auf verteilten Systemen. Die Verteilung erfolgt halbautomatisiert.

Hauptkomponenten:

- HDFS (Hadoop distributed file system)
- MapReduce Implementation

Einige Erweiterungen (+Hauptfunktion):

- Hive (SQL-artige Abfragesprache HiveQL)
- Pig (Skriptsprache, um MapReduce overhead zu verringern)

HiveQl select statement

```
SELECT [ALL | DISTINCT] select_expr, select_expr, ...
FROM table_reference
[WHERE where_condition]
[GROUP BY col_list]
[CLUSTER BY col_list
  | [DISTRIBUTE BY col_list] [SORT BY col_list]
]
[LIMIT number]
```

Pig (Latin)

Pig:

- Pig environment
- Pig Latin

Pig Latin:

- Load $\Rightarrow$ Transform $\Rightarrow$ Dump or Store
- Ausführung per Konsole, Interpreter oder eingebettet in Java Programme
- Was passiert: Script wird per MapReduce auf den Cluster verteilt und ausgeführt
- (Angeblich) komplett, aber UDFs (User defined functions) in Java, Javascript, Python, Ruby möglich

Plan für die nächsten Wochen:

Plan

- Konflikt beseitigen
- Thema/Titel konkretisieren
- Kenntnisse erweitern
- Exposé endlich fertig stellen
- Anmelden
- Literatursuche

References

[AKX+13] Vassil Alexandrov, Jayesh Krishna, Xiabing Xu, Tim Tautges, Iulian Grindeanu, Rob Latham, Kara Peterson, Pavel Bochev, Mary Haley, David Brown, Richard Brownrigg, Dennis Shea, Wei Huang, and Don Middleton.
Parncl and pargal: Data-parallel tools for postprocessing of large-scale earth science data.
In *International Conference on Computational Science, ICCS 2013*. Elsevier, 2013.

[Dur] Jonathan Dursi.
Hpc is dying, and mpi is killing it.
http://www.dursi.ca/hpc-is-dying-and-mpi-is-killing-it/.

[Kun13] Julian Martin Kunkel.
Simulation of Parallel Programs on Application and System Level.
PhD thesis, Universität Hamburg, Von-Melle-Park 3, 20146 Hamburg, 2013.

[Lud14] Prof. Thomas Ludwig.
Hochleistungsrechnen.
Vorlesung, 2014.