8. November 2010

(Advanced) Cloud Computing

Teamprojekt & Projekt

Veranstalter: Prof. Dr. Georg Lausen

Betreuer: Alexander Schätzle,
Martin Przjyaciel-Zablocki,
Thomas Hornung

dbis

Organisation

- **Zeit und Ort:**
  - Montag 14–17 Uhr (c.t.)
  - Raum: SR 00 007 (MMR), Geb. 106
  - Keine weiteren regelmäßigen Treffen

- **Weiterer Ablauf:**
  - Weitere individuelle Termine auf Anfrage
  - 13. Dezember 2010 Zwischenberichte zu den Milestones in getrennten Präsentation (~ 5 min)
  - Gruppe 1: 14 Uhr (c.t)
  - Gruppe 2: 15 Uhr (c.t)

Projektablauf

- ~~Einarbeitungsphase~~
  - Bis Dienstag, 2. November 2010
  - Themenvergabe + Gruppeneinteilung
- ~~Kurzpräsentation~~ heute
  - 8. November 2010
  - Projektvorstellung
  - Eigene Milestones
  - Interne Arbeitsaufteilung
- Implementierungsphase
  - Programmierung & Dokumentation
  - 13. Dezember 2010: Zwischenbericht zu den Milestones (getrennte Präsentationen)
- Abschlusspräsentation
  - 7. Februar 2011
  - Abgabe Projektbericht (14. Februar 2011)

Homepage

- **Neue Materialien**
  - Die neue MapReduce API (diese Folien)
  - MapReduce Überblick/Einführung
  - SVN Anleitung

- **Homepage**
  - `http://dbis.informatik.uni-freiburg.de/index.php?course=WS1011/Projekt/Cloud%20Computing/index.html`

2. MapReduce

» Umstieg auf die neue API

Neue MapReduce API

- Ab Hadoop 0.20 neue API
  - ALT: org.apache.hadoop.**mapred.***
  - NEU: org.apache.hadoop.**mapreduce.***
- Merkmale:
  - Keine neuen Features (noch)
  - Programmierung Java-typischer
  - Codeumfang geringer + übersichtlicher
- CDH 3 Beta
  - Beruht auf Hadoop 0.20.2 + Patches
- **Vorsicht**
  - Änderungen **nicht** kompatibel zueinander
  - Empfehlung: nur neue API verwenden!

Top Level Änderungen

- Methoden können `InterruptedException` als auch `IOException` werfen
- `Configuration` statt `JobConf` Objekt
- Library Klassen wurden nach mapreduce.lib verschoben
  - {input, map, output, partition, reduce}.*

Mapper

- Map Funktion
  - `map (K1 key, V1 value, OutputCollector<K2,V2> output, Reporter reporter)`
  - `map(K1 key, V1 value, Context context)`
- Schließen
  - `Close()`
  - `cleanup(Context context)`
- Ausgabe
  - `Output.collect(K,V)`
  - `Context.write(K,V)`

MapRunnable

- Benutze mapreduce.Mapper
  - `void run (RecordReader<K1,V1> input, OutputCollector<K2,V2> output, Reporter reporter)`
  - `void run (Context context)`

Reducer & Combiner

- Reduce Funktion
  - `void reduce (K2, Iterator<V2> values, OutputCollector<K3,V3> output)`
  - `void reduce (K2, Iterable<V2> values, Context context)`

- Iterationen
  - `while (values.hasNext() { V2 value = values.next(); … }`
  - `for (V2 value: values) { … }`

Submitting Jobs

- `JobConf` + `JobClient` wird durch `Job` ersetzt
  - Job repräsentiert den vollständigen Job statt nur der Konfiguration
  - Job Objekt enthält unterschiedliche Eigenschaften
  - Job Objekt enthält aktuellen Status
  - Warten auf Job bis er ausgeführt wurde möglich

- Job Konstruktor
  - `job = new JobConf(conf, MyMapper.class)`
    `job.setJobName(„job name“)`
  - `job = new Job(conf, „job name“)`
  - `job.setJarByClass(MyMapper.class)`

Submitting Jobs (2)

- Weitere Eigenschaften
  - `Job` hat `getConfiguration`
  - `FileInputFormat` in `mapreduce.lib.input`
  - `FileOutputFormat` in `mapreduce.lib.output`
- Ausführung
  - `JobClient.runJob(job)`
  - `System.exit(job.waitForCompletion(true)?0:1)`

Submitting Jobs (2)

- Weitere Eigenschaften
  - `Job` hat `getConfiguration`
  - `FileInputFormat` in `mapreduce.lib.input`
  - `FileOutputFormat` in `mapreduce.lib.output`
- Ausführung
  - `JobClient.runJob(job)`
  - `System.exit(job.waitForCompletion(true)?0:1)`

Weitere Informationen

- Interessante Links
  - http://www.slideshare.net/sh1mmer/upgrading-to-the-new-map-reduce-api
  - http://hadoop.apache.org/common/docs/r0.20.2/api/
  - http://www.heise.de/developer/artikel/Verarbeiten-grosser-verteilter-Datenmengen-mit-Hadoop-964755.html
  - http://www.heise.de/developer/artikel/Verarbeiten-grosser-verteilter-Datenmengen-mit-Hadoop-ein-Beispiel-1001492.html
  - http://labs.google.com/papers/mapreduce.html
- Buchempfehlung
  - Tom White: Hadoop: The Definitive Guide. *O'Reilly Media*, 06/2009
  - Tom White: Hadoop: The Definitive Guide, Second Edition. *O'Reilly Media*, 09/2010

Agenda

- **Heute**
  - Kurzpräsentation aller Teilnehmer (5-10 Minuten)
  - Inhalt: Projektvorstellung, Milestones und Arbeitsaufteilung
- **Nov, Dez, Jan**
  - Programmierung & Dokumentation
- **Nächstes Treffen**
  - Montag, 13. Dezember 2010 14-17 Uhr (c.t.)
  - Raum: SR 00 007 (MMR), Geb. 106
  - Getrennte Kurzpräsentation (5-10 Minuten)
  - Gruppe 1: 14 Uhr (c.t)
  - Gruppe 2: 15 Uhr (c.t)
  - Inhalt: Zwischenbericht