Scribe & Flume. Datentransport in die Cloud

== Übersicht ==
Die Daten werden auf vielen Nodes gesammelt und sollen in die Cloud geschrieben werden. Die Daten sollen schnell in der Cloud landen, müssen aber um die Clientverbindung zum Node nicht zu belasten nicht live geschrieben werden.

Gerade wenn der Remote Server mal nicht erreichbar sein sollte, sollen die Daten erst lokal gepuffert werden, um sie später zum Remoteserver zu kopieren.
Wir werden die Daten also in einem lokalen Service puffern. Diese Daten sollen dann in bestimmten Intervall von einem Collector abgerufen werden.
Häufiger Anwendungsfall ist der Transport in ein Hadoop Cluster wie er auch in unserer Beispielanwendung behandelt wird.

== Scribe ==
easy-coding.de/Attachment/1115…56ea97ad8c6377d1f78f4bcb0
Scribe wird von Facebook entwickelt und auch von Twitter eingesetzt.
Das Protokoll unter Scribe nutzt zwar Thrift, der Transport eigener Thrift Objekte ist jedoch nicht ohne weiteres möglich.
Ein Datenpaket enthält eine Category und eine Message, die beide als String repräsentiert werden.

Schema in Thrift Notation
Quellcode

```
1. struct LogEntry {
2. 1: string category,
3. 2: string message
4. }
```

Für das Monitoring seiner Scribe Komponenten ist man selbst verantwortlich.

== Flume ==
easy-coding.de/Attachment/1114…56ea97ad8c6377d1f78f4bcb0
Flume wird von Cloudera entwickelt und bietet im Vergleich zu Flume viel mehr Schnittstellen.
Neben dem eigentlichen Inhalt werden auch Quellhost, und beliebige Felder übertragen. Dies ermöglicht leicht flexiblere Anwendungen als Scribe.

Schema in Thrift Notation
Quellcode

```
1. struct ThriftFlumeEvent {
2. 1: Timestamp timestamp,
3. 2: Priority priority,
4. 3: binary body,
5. 4: i64 nanos,
6. 5: string host,
7. 6: map<string,binary> fields
8. }
```

Flume bietet sehr viele Schnittstellen und kann seine Daten auch selbstständig aus anderen Logfiles gewinnen.
Der große Vorteil von Flume ist das Monitoring und die zentrale Konfiguration aller Flume Nodes über den Flume Master.

== Beispielanwendung ==
Im Rahmen dieses Wiki Artikels wird noch ein komplexes Beispiel über das Zusammenspiel von Thrift, Protocol Buffers, Scribe, Flume, Hadoop und Pig erstellt.

Im Beispiel haben wir uns für Flume zu nutzen. Das Datum maskieren wir im Ordnernamen. Als Dateinamen Prefix verwenden wir eine Kategorie wie in Scribe.
* rpcSource(1464)
* collectorSink("hdfs://localhost/tmp/%Y-%m-%d/%H00/", "%{cat}-")