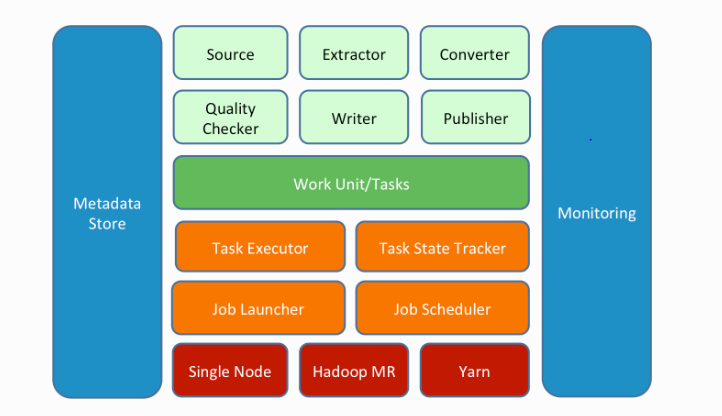
**Gobblin**

* Data integration framework that simplifies common aspects of big data integration such as data ingestion, replication, organization and lifecycle.
* It can be used for streaming and batch data ecosystems.
* It is used for data ingestion from variety of sources(e.g., databases, rest APIs, FTP/SFTP servers, filers, etc.) onto Hadoop.

****

* Highly scalable.(Scheduler, Task Manager)
* Gobblin currently supports two deployment modes: the standalone mode on a single node and the Hadoop MapReduce mode on a Hadoop cluster.
* Its Hadoop integrator.
* Its better if we use it for hdfs(quora).
* It can push data to s3.

**Paper by Linked In**

**Supports heterogeneous data.**

**Doubt===============**

Data is constantly obtained and written into online data storage systems or streaming systems, including Espresso [23], Kafka [6], Voldemort [25], Oracle, MySQL, RocksDB [17], as well as external sites, including S3, Salesforce, Google Analytics, etc. Such data include member profiles, connections, posts and many other external and internal activities. These data sources crop daily terabytes worth of data, most of which needs to be loaded into our Hadoop clusters to feed business- or consumeroriented analysis.