HDFS能够可靠的存储日志和来自不同渠道的其他数据，MapReduce程序能够解析多种特定的（adhoc）数据格式

1. Sqoop允许用户将数据从结构化存储器抽取到Hadoop中。
2. Sqoop连接器

Sqoop拥有一个可扩展的框架，使得它能够从（向）任何支持批量数据传输的外部存储系统导入（导出）数据。

1. 文本和二进制文件格式

Sqoop可以将数据导入成集中不同的格式，默认文本格式，除此还有SequenceFile，Avro，Parquet文件。

1. 深入了解数据库导入
2. 导入控制

Sqoop不需要每次都导入整张表，可以指定仅导入表的部分列

1. 导入和一致性

向HDFS导入数据时，保证一致性的最好的本法是在导入时不允许运行任何对表中现有数据进行更新的进程

1. 增量导入

saved job

1. 导入数据与Hive

Hive和Sqoop共同构成了一个强大的服务于分析任务的工具链

假设有sales.log

首先将数据载入Hive

Hive>create table sales(widget\_id int , qty int, street string, ……) row format delimited fields terminated by ‘,’;

Load data local inpath “sales.log” into table sales;

Sqoop能够根据一个关系数据源中的表来生成一个Hive表，既然已经将widgets表中的数据导入到HDFS，那么我们就直接生成相应的Hive表的定义，然后加载保存在HDFS中的数据。

Sqoop create-hive-table –connect jdbc:mysql://localhost/hadoopguide –table widgets –fields-terminated-by ‘,’

Hive> load data inpath “widgets” into table widgets;

总共三个步骤：将数据导入HDFS🡪创建Hive表🡪将HDFS中的数据导入Hive

使用 –hive-import可以直接从数据源将数据载入hive。

1. 深入理解导出功能

Sqoop导出功能架构和导入功能架构非常类似，Sqoop会根据目标表的定义生成一个java类，这个类可以从文本中解析出记录，并能够向表中插入类型合适的值，然后启动一个MapReduce作业，从HDFS中读取数据源文件

1. 导出与事务

进程的秉性特性决定了导出操作往往不是原子操作。Sqoop会生成多个并行执行的任务，分别导出数据的一部分。

由于在导出过程中如果启动导出结果的应用程序会得到不完整的导出结果，未来解决这个问题，Sqoop可以先将数据导出到一个临时阶段表中，然后在导出任务完成前，在一个事务中将临时阶段表的数据全部移动到目标表中，使用—staging-table。使用临时阶段表会降低执行速度。

1. 导出和SequenceFile