**ASSIGNMENT -22.3**

**Problem Statement**:

Explain in brief

* Sequence File Format
* NLine Input Format
* DB Input Format
* DB Output Format

**Solution:**

1. **Sequence file format:**

Essentially there are 3 different formats for SequenceFiles depending on the CompressionType specified. All of them share a common header described below.

**SequenceFile Header**

* version - 3 bytes of magic header SEQ, followed by 1 byte of actual version number (e.g. SEQ4 or SEQ6)
* keyClassName -key class
* valueClassName - value class
* compression - A boolean which specifies if compression is turned on for keys/values in this file.
* blockCompression - A boolean which specifies if block-compression is turned on for keys/values in this file.
* compression codec - CompressionCodec class which is used for compression of keys and/or values (if compression is enabled).
* metadata - SequenceFile.Metadata for this file.
* sync - A sync marker to denote end of the header.

**Uncompressed SequenceFile Format:**

* Header
* Record
* Record length
* Key length
* Key
* Value
* A sync-marker every few 100 kilobytes or so.

**Record-Compressed SequenceFile Format:**

* Header
* Record
* Record length
* Key length
* Key
* Compressed Value
* A sync-marker every few 100 kilobytes or so.

**Block-Compressed SequenceFile Format:**

* Header
* Record Block
* Uncompressed number of records in the block
* Compressed key-lengths block-size
* Compressed key-lengths block
* Compressed keys block-size
* Compressed keys block
* Compressed value-lengths block-size
* Compressed value-lengths block
* Compressed values block-size
* Compressed values block
* A sync-marker every block.

The compressed blocks of key lengths and value lengths consist of the actual lengths of individual keys/values encoded in ZeroCompressedInteger format.

1. **N Line Input Format:**

* NLineInputFormat which splits N lines of input as one split.
* In many "pleasantly" parallel applications, each process/mapper processes the same input file (s), but with computations are controlled by different parameters.(Referred to as "parameter sweeps"). One way to achieve this, is to specify a set of parameters (one set per line) as input in a control file (which is the input path to the map-reduce application, where as the input dataset is specified via a config variable in JobConf.).
* Performing a large number of these from several reduce tasks concurrently can swamp a database. If you want to export a very large volume of data, you may be better off generating the INSERT statements into a text file, and then using a bulk data import tool provided by your database to do the database import.
* The NLineInputFormat can be used in such applications, that splits the input file such that by default, one line is fed as a value to one map task, and key is the offset. i.e. (k,v) is (LongWritable, Text). The location hints will span the whole mapred cluster.

1. **DB Input Format:**

* Apache Hadoop’s strength is that it enables ad-hoc analysis of unstructured or semi-structured data. Relational databases, by contrast, allow for fast queries of very structured data sources. A point of frustration has been the inability to easily query both of these sources at the same time. The DBInputFormat component provided in Hadoop 0.19 finally allows easy import and export of data between Hadoop and many relational databases, allowing relational data to be more easily incorporated into your data processing pipeline.
* DBInputFormat uses JDBC to connect to data sources. Because JDBC is widely implemented, DBInputFormat can work with MySQL, PostgreSQL, and several other database systems. Individual database vendors provide JDBC drivers to allow third-party applications (like Hadoop) to connect to their databases. To start using DBInputFormat to connect to your database, you’ll need to download the appropriate database driver.
* **Reading Tables with DBInputFormat:**

The DBInputFormat is an InputFormat class that allows you to read data from a database. An InputFormat is Hadoop’s formalization of a data source; it can mean files formatted in a particular way, data read from a database, etc. DBInputFormat provides a simple method of scanning entire tables from a database, as well as the means to read from arbitrary SQL queries performed against the database. Most queries are supported, subject to a few limitations.

1. **DB Output Format:**

* DBOutputFormat accepts <key,value> pairs, where key has a type extending DBWritable. Returned RecordWriter writes only the key to the database with a batch SQL query.
* The DBOutputFormat.setOutput() method then defines how the results will be written back to the database. Its three arguments are the JobConf object for the job, a string defining the name of the table to write to, and an array of strings defining the fields of the table to populate. e.g., DBOutputFormat.setOutput(job, "employees", "employee\_id", "name");.
* The DBOutputFormat writes to the database by generating a set of INSERT statements in each reducer. The reducer’s close() method then executes them in a bulk transaction. Performing a large number of these from several reduce tasks concurrently can swamp a database. If you want to export a very large volume of data, you may be better off generating the INSERT statements into a text file, and then using a bulk data import tool provided by your database to do the database import.