Hadoop InputFormat

Hadoop InputFormat: Types of InputFormat in MapReduce

Hadoop InputFormat checks the Input-Specification of the job. InputFormat split the Input file into InputSplit and assign to individual Mapper.

Let's see:

- What is InputFormat in Hadoop MapReduce,
- different methods to get the data to the mapper and
- different types of InputFormat in Hadoop like FileInputFormat in Hadoop, TextInputFormat, KeyValueTextInputFormat, etc.

We will also see what is the default InputFormat in Hadoop.

What is Hadoop InputFormat?

How the input files are split up and read in Hadoop is defined by the InputFormat. A Hadoop InputFormat is the first component in Map-Reduce, it is responsible for creating the input splits and dividing them into records.

Initially, the data for a MapReduce task is stored in input files, and input files typically reside in HDFS. Although these files format is arbitrary, line-based log files and binary format can be used. Using InputFormat we define how these input files are split and read. The InputFormat class is one of the fundamental classes in the Hadoop MapReduce framework which provides the following functionality:

- The files or other objects that should be used for input is selected by the InputFormat.
- InputFormat defines the Data splits, which defines both the size of individual Map tasks and its potential execution server.
- InputFormat defines the RecordReader, which is responsible for reading actual records from the input files.

How we get the data to mapper?

We have 2 methods to get the data to mapper in MapReduce: `getsplits()` and `createRecordReader()`.

Types of InputFormat in MapReduce

FileInputFormat in Hadoop

It is the base class for all file-based InputFormats. Hadoop FileInputFormat specifies input directory where data files are located. When we start a Hadoop job, FileInputFormat is provided with a path containing files to read. FileInputFormat will read all files and divides these files into one or more InputSplits.

TextInputFormat

It is the default InputFormat of MapReduce. TextInputFormat treats each line of each input file as a separate record and performs no parsing. This is useful for unformatted data or line-based records like log files.

- Key – It is the byte offset of the beginning of the line within the file (not whole file just one split), so it will be unique if combined with the file name.

- Value – It is the contents of the line, excluding line terminators.

KeyValueTextInputFormat

It is similar to TextInputFormat as it also treats each line of input as a separate record. While TextInputFormat treats entire line as the value, but the KeyValueTextInputFormat breaks the line itself into key and value by a tab character ('/t'). Here Key is everything up to the tab character while the value is the remaining part of the line after tab character.

SequenceFileInputFormat

Hadoop SequenceFileInputFormat is an InputFormat which reads sequence files. Sequence files are binary files that stores sequences of binary key-value pairs. Sequence files block-compress and provide direct serialization and deserialization of several arbitrary data types (not just text). Here Key & Value both are user-defined.

SequenceFileAsTextInputFormat

Hadoop SequenceFileAsTextInputFormat is another form of SequenceFileInputFormat which converts the sequence file key values to Text objects. By calling 'tostring()' conversion is performed on the keys and values. This InputFormat makes sequence files suitable input for streaming.

SequenceFileAsBinaryInputFormat

Hadoop SequenceFileAsBinaryInputFormat is a SequenceFileInputFormat using which we can extract the sequence file's keys and values as an opaque binary object.

NLineInputFormat

Hadoop NLineInputFormat is another form of TextInputFormat where the keys are byte offset of the line and values are contents of the line. Each mapper receives a variable number of lines of input with TextInputFormat and KeyValueTextInputFormat and the number depends on the size of the split and the length of the lines. And if we want our mapper to receive a fixed number of lines of input, then we use NLineInputFormat.

N is the number of lines of input that each mapper receives. By default (N=1), each mapper receives exactly one line of input. If N=2, then each split contains two lines. One mapper will receive the first two Key-Value pairs and another mapper will receive the second two key-value pairs.

DBInputFormat

Hadoop DBInputFormat is an InputFormat that reads data from a relational database, using JDBC. As it doesn't have portioning capabilities, so we need to careful not to swamp the database from which we are reading too many mappers. So, it is best for loading relatively small datasets, perhaps for joining with large datasets from HDFS using MultipleInputs. Here Key is LongWritables while Value is DBWritables.

Hadoop InputFormat

Hadoop OutputFormat: Types of OutputFormat in MapReduce