Date	Time	Status	Type	Tasks/Appointments	Remarks
26/2	9:00:00	Clear	--	CSIT Orientation - HR / Security Briefing - Settle Logistic	
28/2				Database mindmap -Explored different types DBs including distributed DB, Data Warehouse etc. -Understand the functional and operational requirements of DB -Explored the different Data Models e.g. Hierarchical, network, relational, XML etc. -Basic understanding of DB architecture: External level, Conceptual level, Internal level	To-do: -In-depth research on Data Warehouse and create a mindmap to aid understanding Keywords: Data Models, Relational Databases, Data Independence
28/2				Data Warehouse -Layers of a DW (process flow of a DW from feeding data from sources into DW to presentation of Data -Different approaches of DW: Normalized and Dimensional -Benefits of DW -Design Methodologies: Bottom-up and Top-down -OLAP (On-Line Analytical Processing)	To-do: -Research on Normalized and Dimensional Databases -Research on OLAP Keywords: ETL (Extraction, Transformation, Loading), Normalized Data (1NF 2NF 3NF), Dimensional Data, Star/Snowflake Schema, OLAP
29/2	11:30:00			Courtesy Call with Senior Manager (Winston Chan)	
				Relational Database -Understood main terminologies associated with relational databases (relation, tuples, attributes, superkey, primary key, foreign key, integrity rules -Write up on relational Database	To-do: - Also focus on Relational Database, Column oriented Database - Also readup on SQL and noSQL - Coverage of Data Warehouse is extensive.
				Column-oriented Database -Short write-up of research	- Analysis of application of relational database, column-oriented database to problem sets
				Preliminary reading on -Hadoop -Hbase and Big Table	
1/3/2012				Readings on: -noSQL -Column-oriented database -Document-oriented database	More questions for column-oriented db: (Maybe relevant to Hbase storage structure) What is its query language? SQL like? What are the implementation/usage for column-oriented db? What are the disadvantage of column-oriented db? Not suitable for certain scenario etc.
					Database Filesystem - http://dbfs.sourceforge.net/ How hadoop and technologies differ for such system?
2/3/2012				Column-oriented Database -Completion of write-up Introduction to Hadoop -Yahoo tutorial	
5/3/2012				Exploration of hadoop program -interaction with Ubuntu Linux OS -setting up of single-node server -module 3 of yahoo tutorial	samba
8/3/2012				M/R program for word count completed M/R program for reverse indexing successful	
13/3/2012				Understanding of M/R data flow Reading on custom data types for M/R	MapReduce Data Flow diagram up
14/3/2012				Reading on custom FileInputFormat, RecordReader, FileOutputFormat, RecordWriter Insanely difficult to understand	
				Possible exploration of MapR software	
					Do spend some effort to journal your research/hand-ons/reading progresses. E.g. like which web tutorial/video you covered for that day, your findings and rating/comments etc. It will be a good reference for the tonnes of material you will be covering by the end of the internship. It will be your knowledge database to take away. So do feel free to arrange it in any way you want. It is a chore definitely but it is also a very useful tool. So a bit more of discipline will help :) Just to share, i keep a similar simple work journal as well. Is very useful especially at time to time we need to do work review.
15/3/2012				managed to complete yahoo tutorial on ObjPosInputFormat, ObjectPositionReader and XMLOutputFormat managed to follow the tutorial and set up custom reporter metrics using Reporter.incrCounter() introduction to DistributedCache class to send data files that name nodes need to them. http://developer.yahoo.com/hadoop/tutorial/module5.html	Java BufferedStream methods used for recordreader need to explore and learn more about it to build inputformats customized for our needs
16/3/2012				Exploring Pig for hadoop - SQL type layer on top of hadoop http://pig.apache.org/docs/r0.9.2/start.html	cant have any hands-on as pig is not installed on the vmware image
19/3/12				Back from weekend, back to hadoop will attempt to code an input format for separating text by '$' to learn more about Java Manage to use FileInputStream to read text files and separate phrases by '$' at the byte-level, but incorporating it together with hadoop will be a different story	*Request for thumbdrive please to transfer Pig source files to dev machine exploration on Pig

20/3/12				Pig basics: LOAD, DUMP, STORE, FOREACH...GENERATE, FILTER...BY, SPLIT...INTO...IF, JOIN...BY, GROUP...BY, ORDER...BY, LIMIT, UNION, DISTINCT http://pig.apache.org/docs/r0.9.2/basic.html	
23/3/12				Reading Hadoop: The Definitive Guide	Hadoop documentation on what i've learnt
26/3/12				Organizing Hadoop documentation for clearer presentation Consolidation of knowledge on Pig	Preliminary exploration on Cassandra and Hbase
27/3/12				Preliminary exploration of HBase	Hands-on with development machine
30/3/12					Sunjing: The first remote Ubuntu node is up with remote desktop and smbmount enabled. IP: 10.2.41.101
4/4/2012					Sunjing: Transferred in the UbuntuRepos\pool\restricted and \universe folder. Still have abt 2GB+ outstanding files. To rerun dpkg-scanpackages and upload the difference into terastation.
25/4/2012				Documentation for Hadoop single-node set up	
				Documentation for Hadoop distributed set up	
				Documentation for Zookeeper replicated set up	
				Documentation for HBase distributed set up	
				Successfully set up HDFS, ZK, HBASE on 3 nodes	
30/4				Trying to tweak TextInputFormat and LineRecordReader, such that inputformat will read each line in the file to retrieve the file name and also the current path of the index.txt file	
				This is based on the assumption that index.txt file is in the same HDFS directory with the rest of the news article html files	
				Having the path as output will make the MR more flexible and less hard-coded	
				A particular method available from Hadoop's Path class is toUri(), which converts a Path object into the Java URI object. This opens up more possibilities	
				Right now what I'm doing is directly convert the Path object to the string, and the Path object holds the information to the input file defined in args[0]	
				so for now printing Path.toString() will give me hdfs://master:54310/user/hduser/news/2006-01/index.txt	
				using the method .getParent, Path.getParent().toString() will get me to hdfs://master:54310/user/hduser/news/2006-01	
				which is where all the article html files reside.	
2/5/2012				Manage to retrieve contents for articles html files using BufferInputStream.read() to find out buffer size, and Text.append()	
				Trying to tweak InputFormat such that file is read and contents retrieved first	
3/5/2012				Success! Tweaked InputFormat such that the input to the mapper is (0, <html contents>)	
				Next step is to tweak it such that input to mapper is (<article id>,<html contents>)	
4/5/2012				Implemented my own ArticleID class	
8/5/2012				Connected HBase to HDFS, distributed set-up. need to document (note on adding hbase classpath to hadoop)	
9/5/2012				Able to use TableOutputFormat to write to HBase, must remeber that Table and CF needs to exist/create beforehand	
				Try to use HFileOutputFormat and then completebulkloader (failed initially with an error saying split found when grouping hfiles)	
11/5/2012				calling .add() on put object to insert value into different columns (same row key)	
				added logic to HtmlParser and catching possible exceptions (incomplete)	
22/5/2012				Changed rowkey to be filename (rather than a composite of filename and article title, under a new object called ArticleID)	
				Added documentation (following Java Code Conventions) to classes and methods	
				Improved ease of use by allowing user to specify table and column family to write to at the command line	
				Aim to reduce hard coding (by offering more options at command line	
14/6/12				Successfully created Record and Row class, Row class has methods to add records/rows to existing rows (merge rows) so that a single row is made up of a single row-id	
25/6/12				Documentation stage begins	