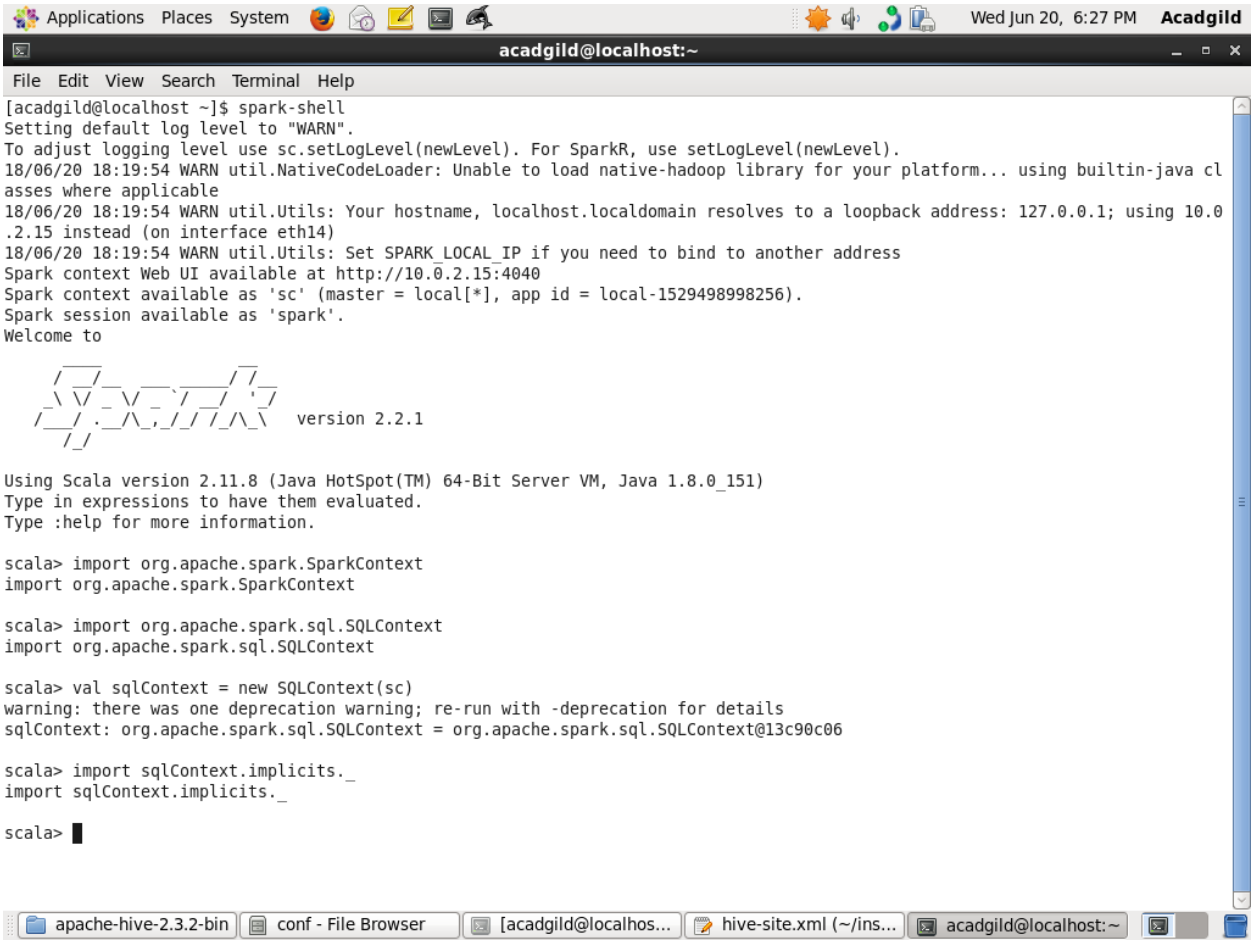
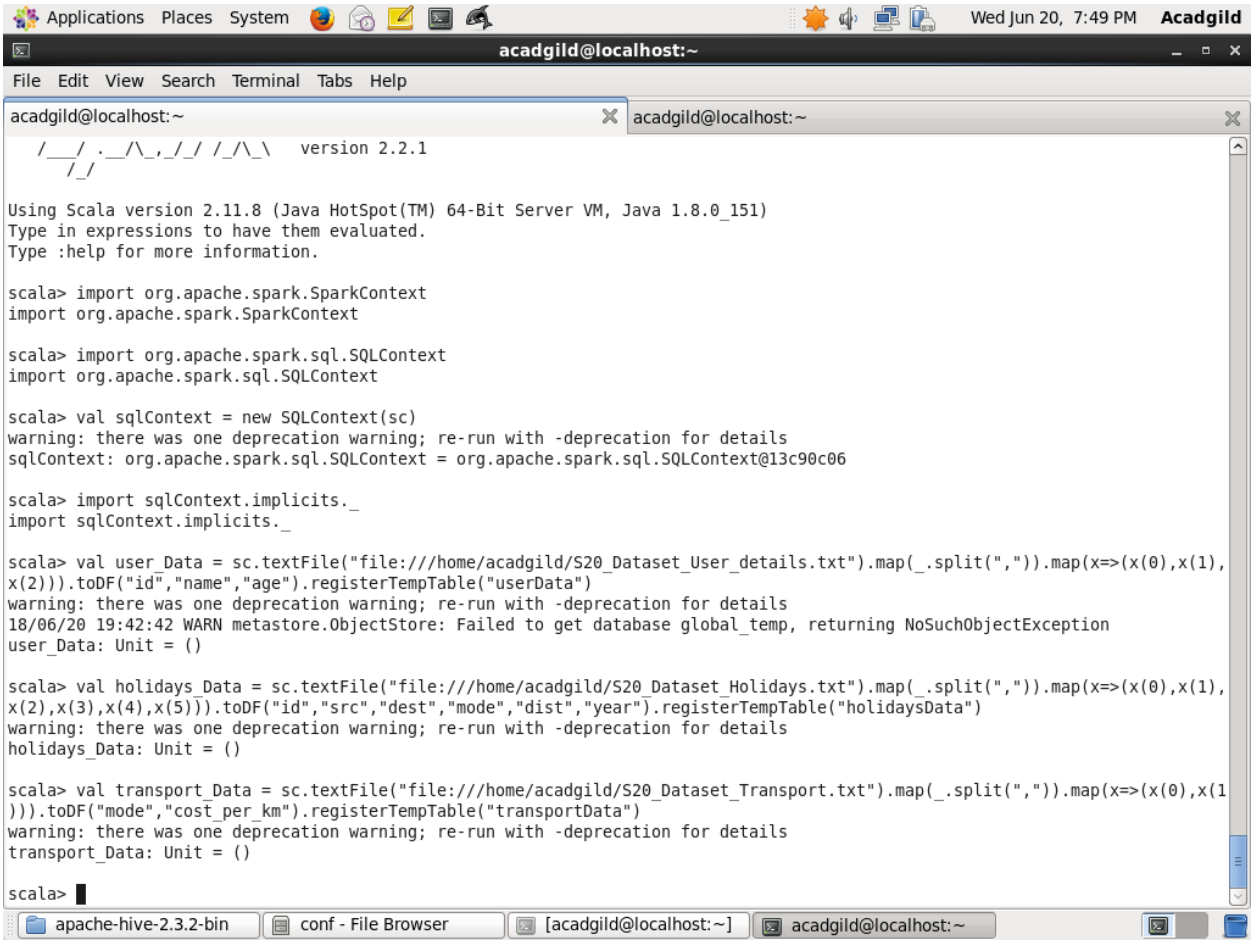
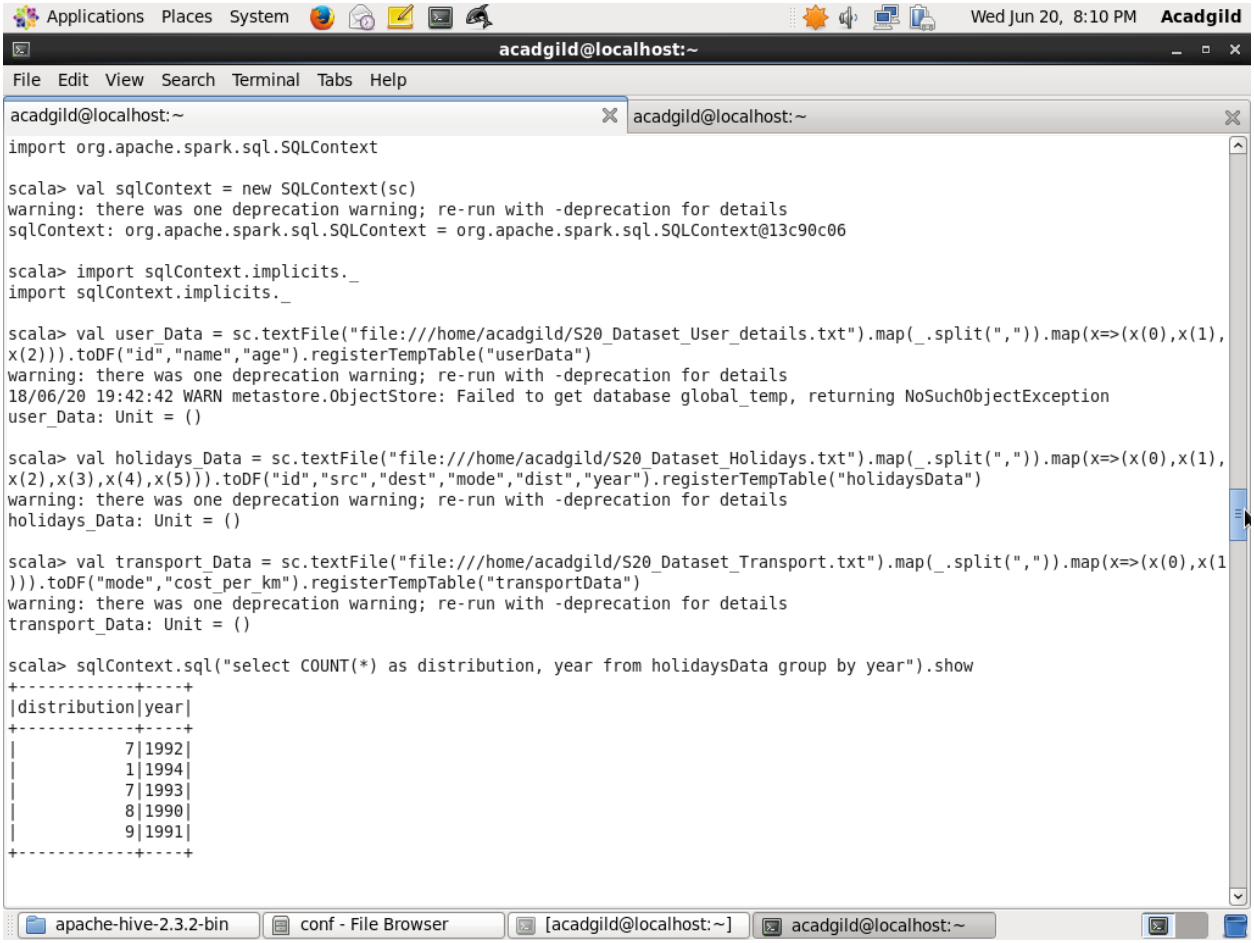
Performing following steps to use sql context in sparkshell



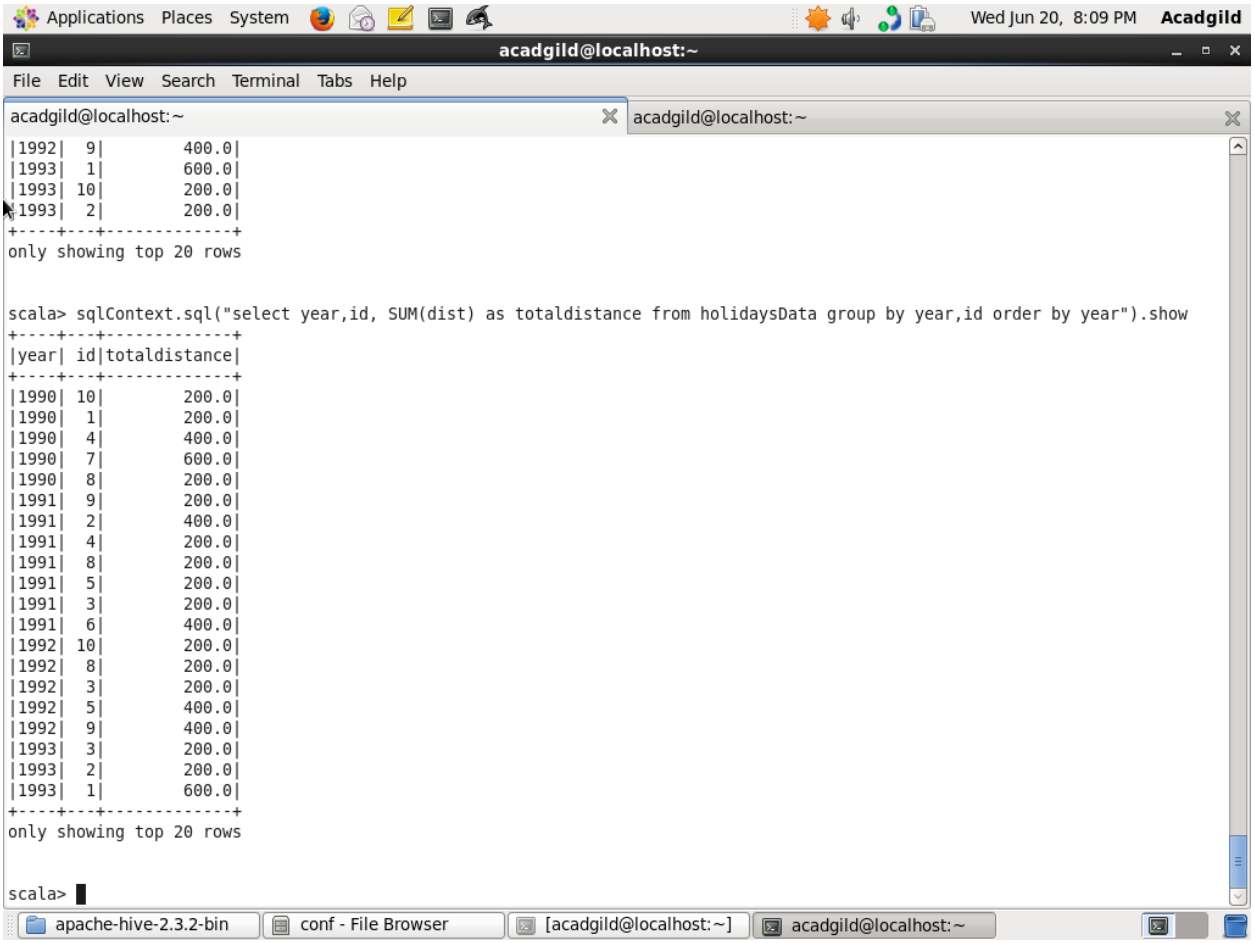
Loading data as dataframe with column name and registering as temp table



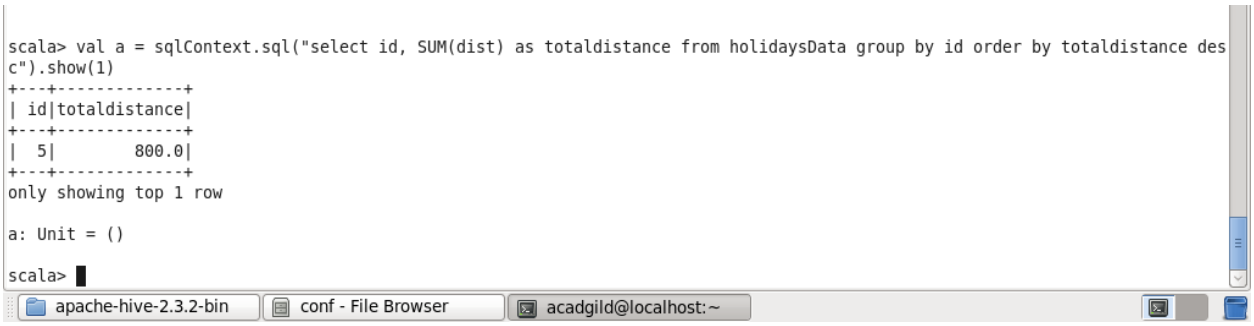
Finding distribution of total number of air travellers per year



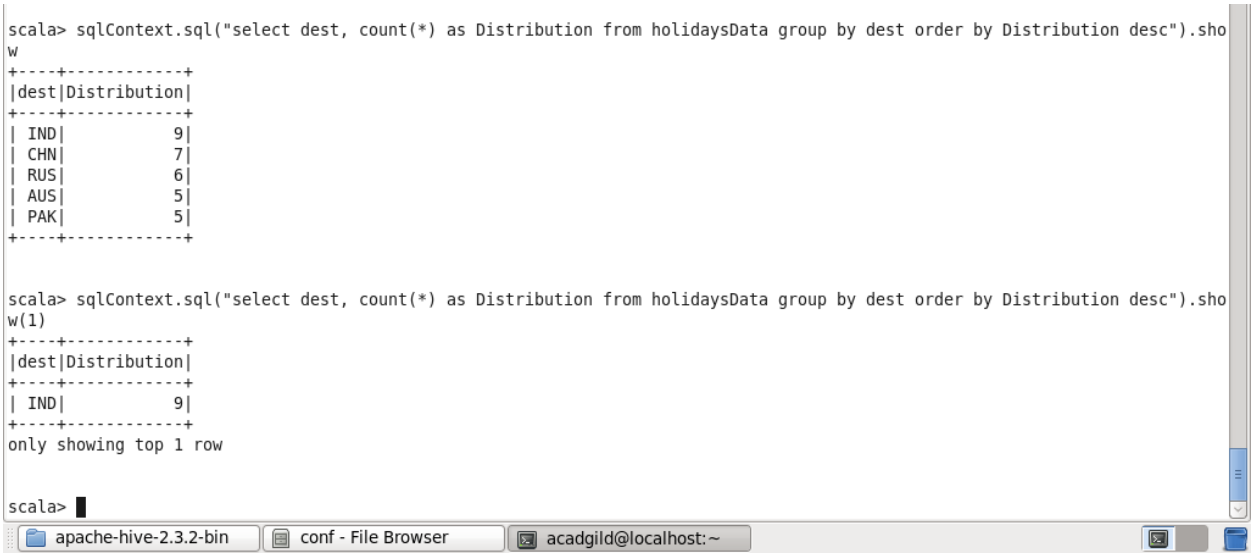
Total distance covered by each user every year



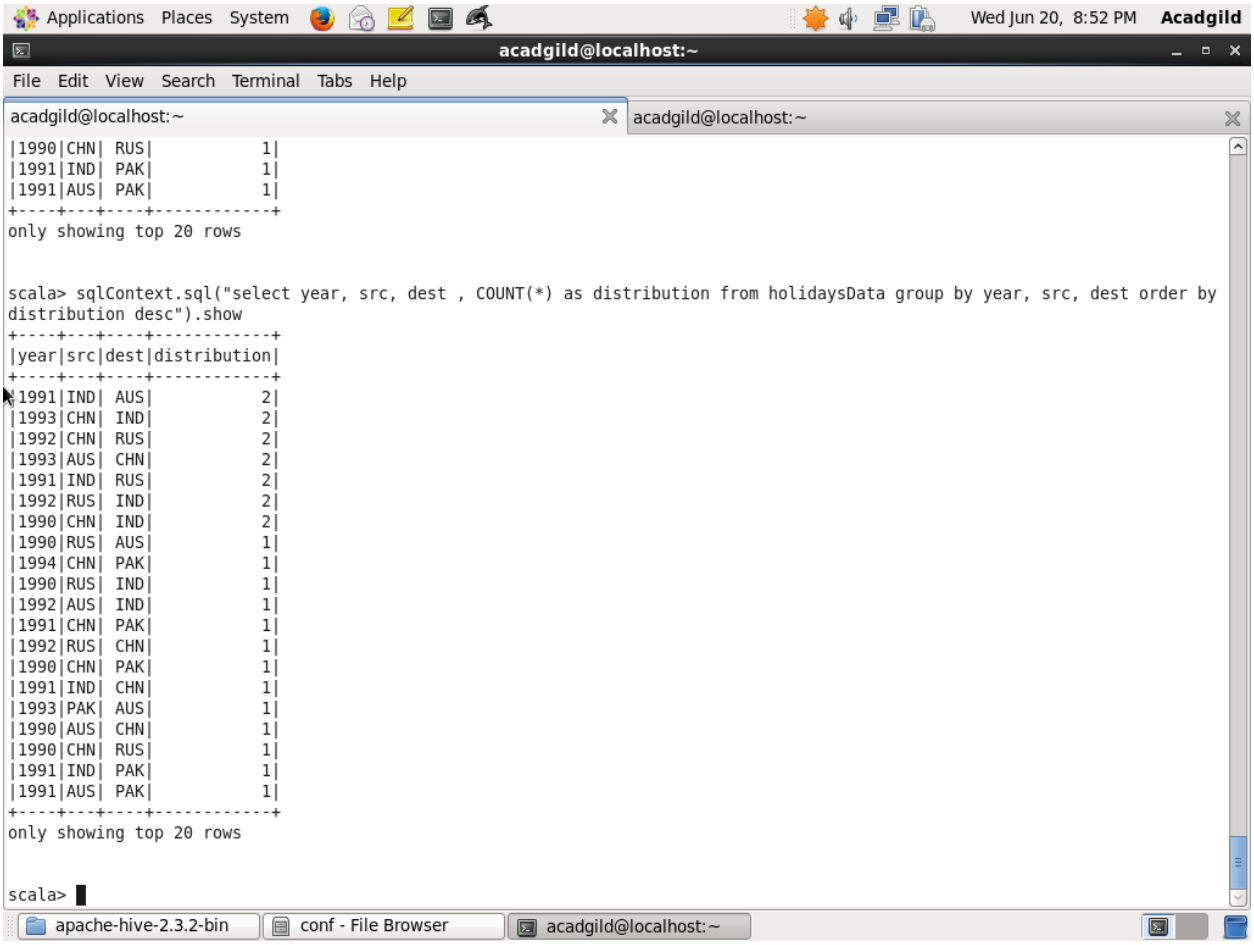
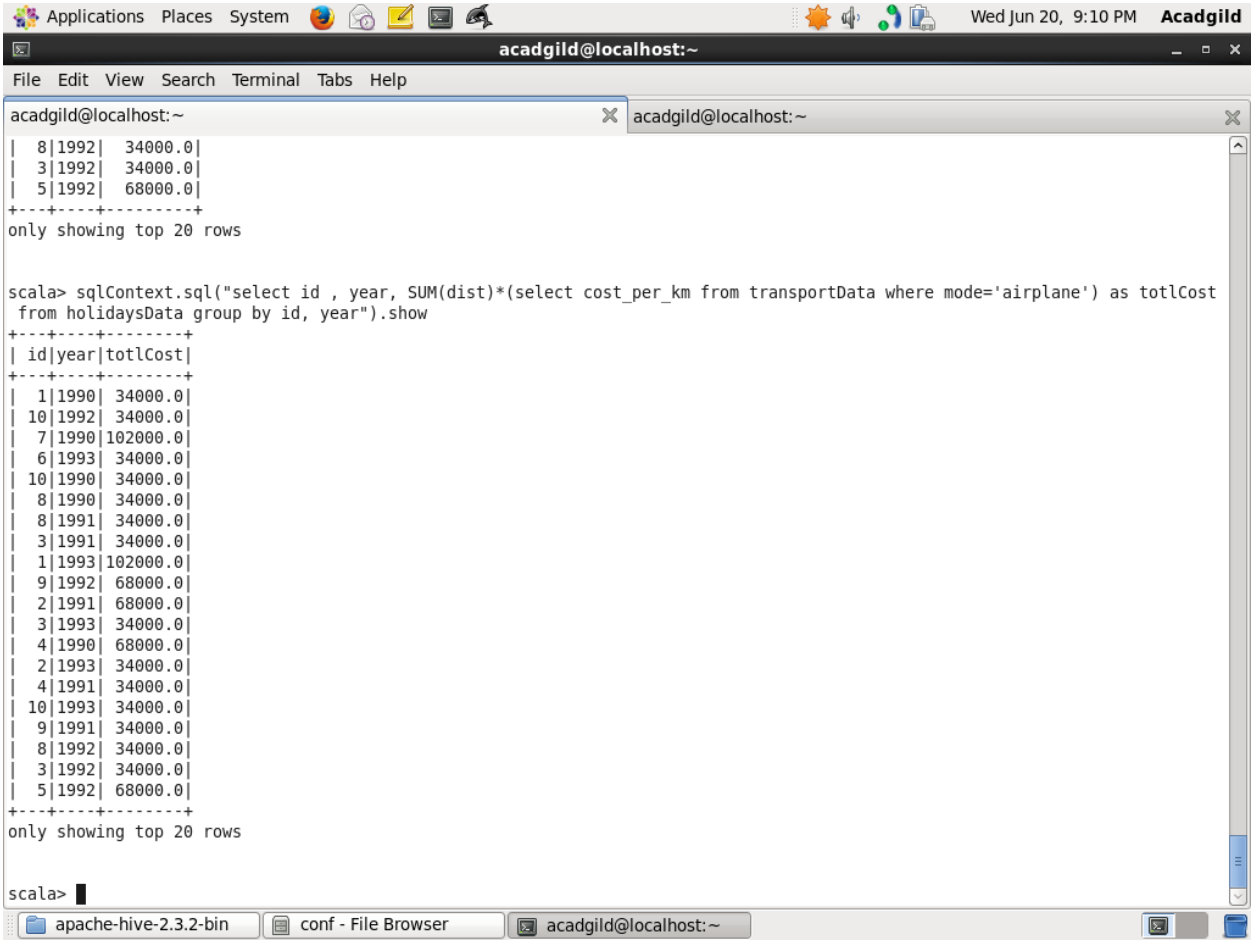
Finding user who has travelled the most distance :



Finding most preferred distance for all users:



Route generating most revenue per year:


Applications Places System
Wed Jun 20, 8:52 PM Acadgild
acadgild@localhost:~
File Edit View Search Terminal Tabs Help
acadgild@localhost:~
acadgild@localhost:~
1990	CHN	RUS	1
1991	IND	PAK	1
1991	AUS	PAK	1
+----+---+----+-----------+
only showing top 20 rows
scala> sqlContext.sql("select year, src, dest , COUNT(*) as distribution from holidaysData group by year, src, dest order by distribution desc").show
+----+---+----+-----------+
|year|src|dest|distribution|
+----+---+----+-----------+
1991	IND	AUS	2
1993	CHN	IND	2
1992	CHN	RUS	2
1993	AUS	CHN	2
1991	IND	RUS	2
1992	RUS	IND	2
1990	CHN	IND	2
1990	RUS	AUS	1
1994	CHN	PAK	1
1990	RUS	IND	1
1992	AUS	IND	1
1991	CHN	PAK	1
1992	RUS	CHN	1
1990	CHN	PAK	1
1991	IND	CHN	1
1993	PAK	AUS	1
1990	AUS	CHN	1
1990	CHN	RUS	1
1991	IND	PAK	1
1991	AUS	PAK	1
+----+---+----+-----------+
only showing top 20 rows
scala>
apache-hive-2.3.2-bin
conf - File Browser
acadgild@localhost:~

Amount spent by every user on air travel per year :