Cheatography

DL Cheat Sheet

by woobidoobi via cheatography.com/99373/cs/21014/

Session

Interpreter	%jdbc(-hive)	%livy2.pyspark
Create Session		"from pyspark_llap import HiveWarehouseSession hive = HiveWarehouseSession.session(spark).build()"
Connect to database	use ampb_sandbox	"hive.setDatabase(""dl_prod_sandbox_agi"") "
List Databases		hive.execute("show databases").show(truncate=False)
List Tables	show tables	hive.execute("show tables").show(truncate=False)

Data wrangling

function	sql equivalient	pyspark df
Selecting columns	select arbetsgivaravgift_organisationsnummer_id as orgnr, arbetsgivaravgift_redovisad_period as period, arbetsgivaravgift_inbetalat_belopp as inbetalt_belopp from df_agavgift	df_agavgift.select(col("arbetsgivaravgift_organisationsnummer_id").alias("orgnr"), col("arbetsgivaravgift_redovisad_period").alias("period"), col("arbetsgivaravgift_inbetalat_belopp").alias("inbetalat_belopp"))

Livy2.pyspark

Transfer data from hive to pyspark dataframe	pyspark_df = hive.table("hive_table")
Transfer data from pyspark to hive	pyspark_df.registerTempTable("hive_table")
Remove a temp table	hive_table.drop()

Combining data

inner join	df1.join(df2, df1.name == df2.name)
left join	df1.join(df2, df1.name == df2.name,how='left')
right join	ta.join(tb, ta.name == tb.name,how='right')
joining on multiple columns	df1.join(df2, df1.name == df2.name,how='right')

in pyspark you need to start by setting an alias for the tables that you want to join
df1 = TableA.alias('df1')
df2 = TableB.alias('df2')

Useful links:
http://www.learnbymarketing.com/1100/pyspark-joins-by-example/
http://www.learnbymarketing.com/618/pyspark-rdd-basics-examples/