SQL TO PIG
Cheat Sheet

We know that lots of people come to Apache Pig from a relational database background, so we compiled this handy translation from SQL concepts to their Pig equivalents.

SQL CONCEPT	SQL	PIG

SELECT	SELECT column_name,column_name
	FROM table_name;
		FOREACH alias GENERATE column_name,
		column_name;

SELECT *	SELECT * FROM table_name;	FOREACH alias GENERATE *;

DISTINCT	SELECT DISTINCT column_name,column_name
	FROM table_name;
		DISTINCT(FOREACH alias
		GENERATE column_name, column_name);

WHERE	SELECT column_name,column_name
	FROM table_name
	WHERE column_name operator value;
		FOREACH (FILTER alias BY column_name
		operator value)
		GENERATE column_name, column_name;

AND/OR	... WHERE (column_name operator value1 AND column_name operator value2) OR column_name operator value3;	FILTER alias BY (column_name operator value1 AND column_name operator value2) OR column_name operator value3;
ORDER BY	... ORDER BY column_name ASC|DESC, column_name ASC|DESC;	ORDER alias BY column_name ASC|DESC, column_name ASC|DESC;

TOP/LIMIT	SELECT TOP number column_name
	FROM table_name ORDER BY column_name ASC|DESC;
	
	SELECT column_name FROM table_name
	ORDER BY column_name ASC|DESC LIMIT number;
		TOP(number, column_index, alias);
		
		FOREACH (GROUP alias BY column_name)
		GENERATE LIMIT alias number;

GROUP BY	SELECT function(column_name)
	FROM table GROUP BY column_name;
		FOREACH (GROUP alias BY column_name)
		GENERATE function(alias.column_name);

LIKE	... WHERE column_name LIKE pattern;	FILTER alias BY REGEX_EXTRACT(column_name, pattern, 1) IS NOT NULL;
IN	... WHERE column_name IN (value1,value2,...);	FILTER alias BY column_name IN (value1, value2,...);

SQL CONCEPT	SQL	PIG

JOIN	SELECT column_name(s)
	FROM table1
	JOIN table2
	ON table1.column_name=table2.column_name;
		FOREACH (JOIN alias1 BY column_name,
		alias2 BY column_name)
		GENERATE column_name(s);

LEFT/RIGHT/FULL OUTER JOIN	SELECT column_name(s)
	FROM table1
	LEFT|RIGHT|FULL OUTER JOIN table2
	ON table1.column_name=table2.column_name;
		FOREACH (JOIN alias1 BY column_name
		LEFT|RIGHT|FULL, alias2 BY column_name)
		GENERATE column_name(s);

UNION ALL	SELECT column_name(s) FROM table1
	UNION ALL
	SELECT column_name(s) FROM table2;
		UNION alias1, alias2;

AVG	SELECT AVG(column_name) FROM table_name;	FOREACH (GROUP alias ALL) GENERATE AVG(alias.column_name);
COUNT	SELECT COUNT(column_name) FROM table_name;	FOREACH (GROUP alias ALL) GENERATE COUNT(alias);

COUNT DISTINCT	SELECT COUNT(DISTINCT column_name) FROM table_name;	FOREACH alias {
		unique_column = DISTINCT column_name;
		GENERATE COUNT(unique_column);
		};

MAX	SELECT MAX(column_name) FROM table_name;	FOREACH (GROUP alias ALL) GENERATE MAX(alias.column_name);
MIN	SELECT MIN(column_name) FROM table_name;	FOREACH (GROUP alias ALL) GENERATE MIN(alias.column_name);
SUM	SELECT SUM(column_name) FROM table_name;	FOREACH (GROUP alias ALL) GENERATE SUM(alias.column_name);
HAVING	... HAVING aggregate_function(column_name) operator value;	FILTER alias BY aggregate_function(column_name) operator value;
UCASE/UPPER	SELECT UCASE(column_name) FROM table_name;	FOREACH alias GENERATE UPPER(column_name);
LCASE/LOWER	SELECT LCASE(column_name) FROM table_name;	FOREACH alias GENERATE LOWER(column_name);

SUBSTRING	SELECT SUBSTRING(column_name,start,length)
	AS some_name FROM table_name;
		FOREACH alias GENERATE SUBSTRING(column_name,
		start, start+length) as some_name;

LEN	SELECT LEN(column_name) FROM table_name;	FOREACH alias GENERATE SIZE(column_name);
ROUND	SELECT ROUND(column_name,0) FROM table_name;	FOREACH alias GENERATE ROUND(column_name);

For more tips, download the full Pig Cheat Sheet: bit.ly/pigcheat