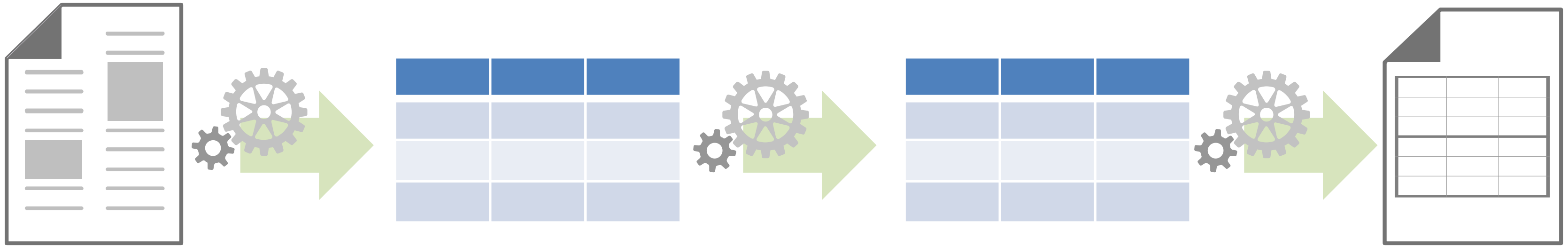
Beyond Hive – Pig and Python

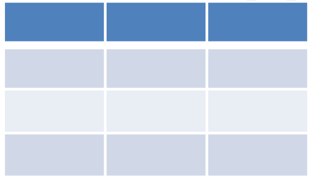
Microsoft

What is Pig?



- Pig performs a series of transformations to data *relations* based on *Pig Latin* statements
- Relations are loaded using *schema on read* semantics to project table structure at runtime
- You can run Pig Latin statements interactively in the *Grunt* shell, or save a script file and run them as a batch

- A *relation* is an *outer bag*
  - A *bag* is a collection of *tuples*
  - A *tuple* is an ordered set of *fields*
  - A *field* is a data item
- A *field* can contain an *inner bag*
- A *bag* can contain *tuples* with non-matching schema



{(a, 1)
(b, 2)
(c, 3) }
(d, {(4, 5), (6,7)}) }
(e)
(f, 8, 9) }

What kinds of things can I do with Pig?

```
2013-06-01,12
2013-06-01,14
2013-06-01,16
2013-06-02,9
2013-06-02,12
2013-06-02,9
...
```

```
-- Load comma-delimited source data
Readings = LOAD '/weather/data.txt' USING PigStorage(',') AS (date:chararray, temp:long);
-- Group the tuples by date
GroupedReadings = GROUP Readings BY date;
-- Get the average temp value for each date grouping
GroupedAvgs = FOREACH GroupedReadings GENERATE group, AVG(Readings.temp) AS avgtemp;
-- Ungroup the dates with the average temp
AvgWeather = FOREACH GroupedAvgs GENERATE FLATTEN(group) as date, avgtemp;
-- Sort the results by date
SortedResults = ORDER AvgWeather BY date ASC;
-- Save the results in the /weather/summary folder
STORE SortedResults INTO '/weather/summary';
```

```
2013-06-01  14.00
2013-06-02  10.00
```

Common Pig Latin Operations

- LOAD
- FILTER
- FOR EACH … GENERATE
- ORDER
- JOIN
- GROUP
- FLATTEN
- LIMIT
- DUMP
- STORE

- Pig generates Map and Reduce operations from Pig Latin
- Jobs are generated on:
  - DUMP
  - STORE

```
Readings = LOAD '/weather/data.txt' USING PigStorage(',') AS (date, temp:long);
GroupedReadings = GROUP Readings BY date;
GroupedAvgs = FOREACH GroupedReadings GENERATE group, AVG(Readings.temp) AS avgtemp;
AvgWeather = FOREACH GroupedAvgs GENERATE FLATTEN(group) as date, avgtemp;
SortedResults = ORDER AvgWeather BY date ASC;
STORE SortedResults INTO '/weather/summary';
```

Job generated here

How do I run a Pig script?

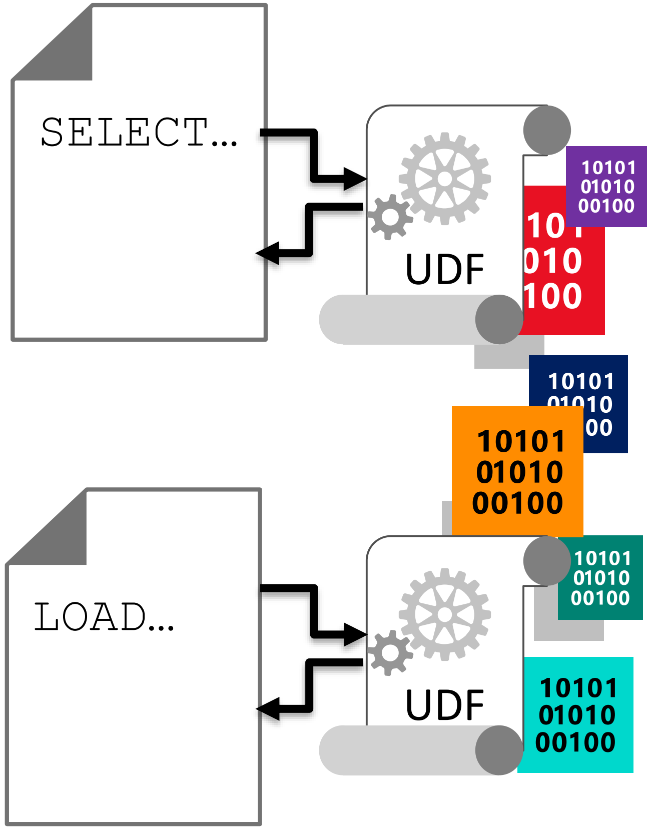
1. Save a Pig Latin script file
2. Run the script using Pig

```
pig wasb:///scripts/myscript.pig
```

3. Consume the results using any Azure storage client
   - For example, Excel or Power BI
   - Default output does not include schema – just data

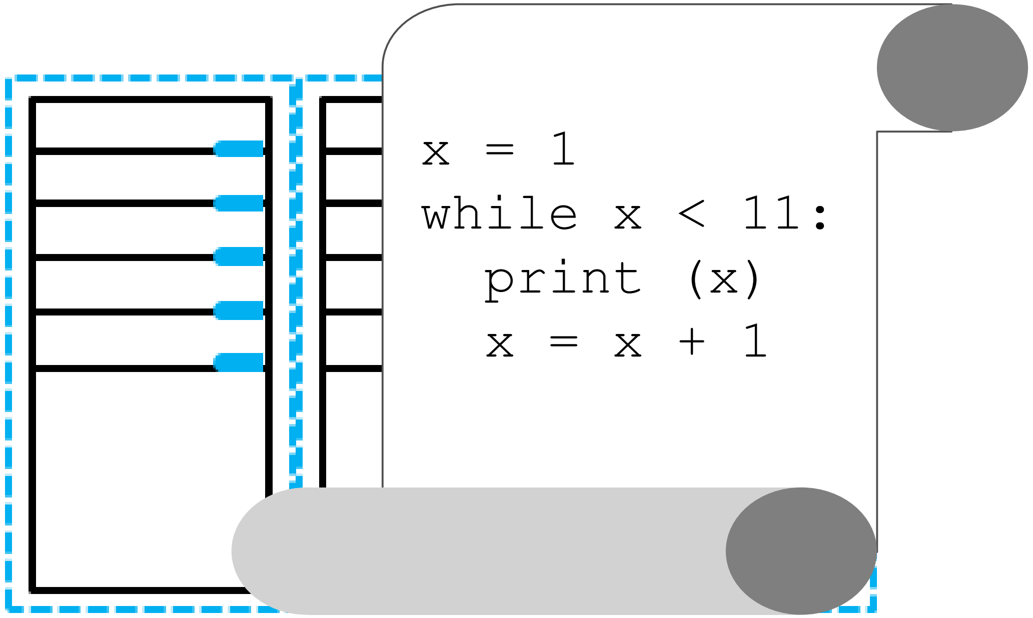
What are UDFs?

- User-Defined Functions (UDFs) extend the capabilities of Hive and Pig
- Simpler than writing custom MapReduce components
- Can be implemented using many languages, for example:
  - Java
  - C#
  - Python



Python is a (relatively) simple scripting language – ideal for UDFs

- Intuitive syntax
- Dynamic typing
- Interpreted execution



```
x = 1
while x < 11:
  print (x)
  x = x + 1
```

Python is pre-installed on HDInsight clusters

- Python 2.7 supports *streaming* from Hive
- Jython (a Java implementation of Python) has native support in Pig

How do I use a Python UDF in Pig?

Pig natively supports Jython

- Define the output schema as a Pig bag
- Declare a Python function that receives an input parameter from Pig
- Return results as fields based on the output schema

```
@outputSchema("result: {(a:chararray, b:int)}")
Def myfunction(i):
  ...

  return a, b
```

Use the Pig FOREACH…GENERATE statement to invoke a UDF

```
REGISTER 'wasb:///scripts/myscript.py' using jython as myscript;

src = LOAD '/data/source' AS (row:chararray);

res = FOREACH src GENERATE myscript.myfunction(row);
```

How do I use a Python UDF in Hive?

Hive exchanges data with Python using a *streaming* technique

- Rows from Hive are passed to Python through STDIN
- Processed rows from Python are passed to Hive through STDOUT

```
line = sys.stdin.readline()

...

print processed_row
```

Use the Hive TRANSFORM statement to invoke a UDF

```
add file wasb:///scripts/myscript.py;

SELECT TRANSFORM (col1, col2, col3)
  USING 'python myscript.py'
  AS(col1 string, col2 int, col3 string)
FROM mytable
ORDER BY col1;
```

Microsoft