

“Big Data Processing – *on the cheap...*”



Joe Hummel, PhD

Visiting Researcher: U. of California, Irvine
Adjunct Professor: U. of Illinois, Chicago &
Loyola U., Chicago

Materials: <http://www.joehummel.net/downloads.html>
Email: joe@joehummel.net

Agenda

- ▶ Three inexpensive ways to process big data
 1. **PowerPivot** plugin for Microsoft Excel
 2. **LINQ**-based programming approach in .NET
 3. **Hadoop** map-reduce framework
- ▶ Discuss trade-offs
- ▶ Demos, demos, demos

How big is big?

- ▶ What is your definition of big data?
 - MBs?
 - GBs?
 - TBs?
 - PBs?

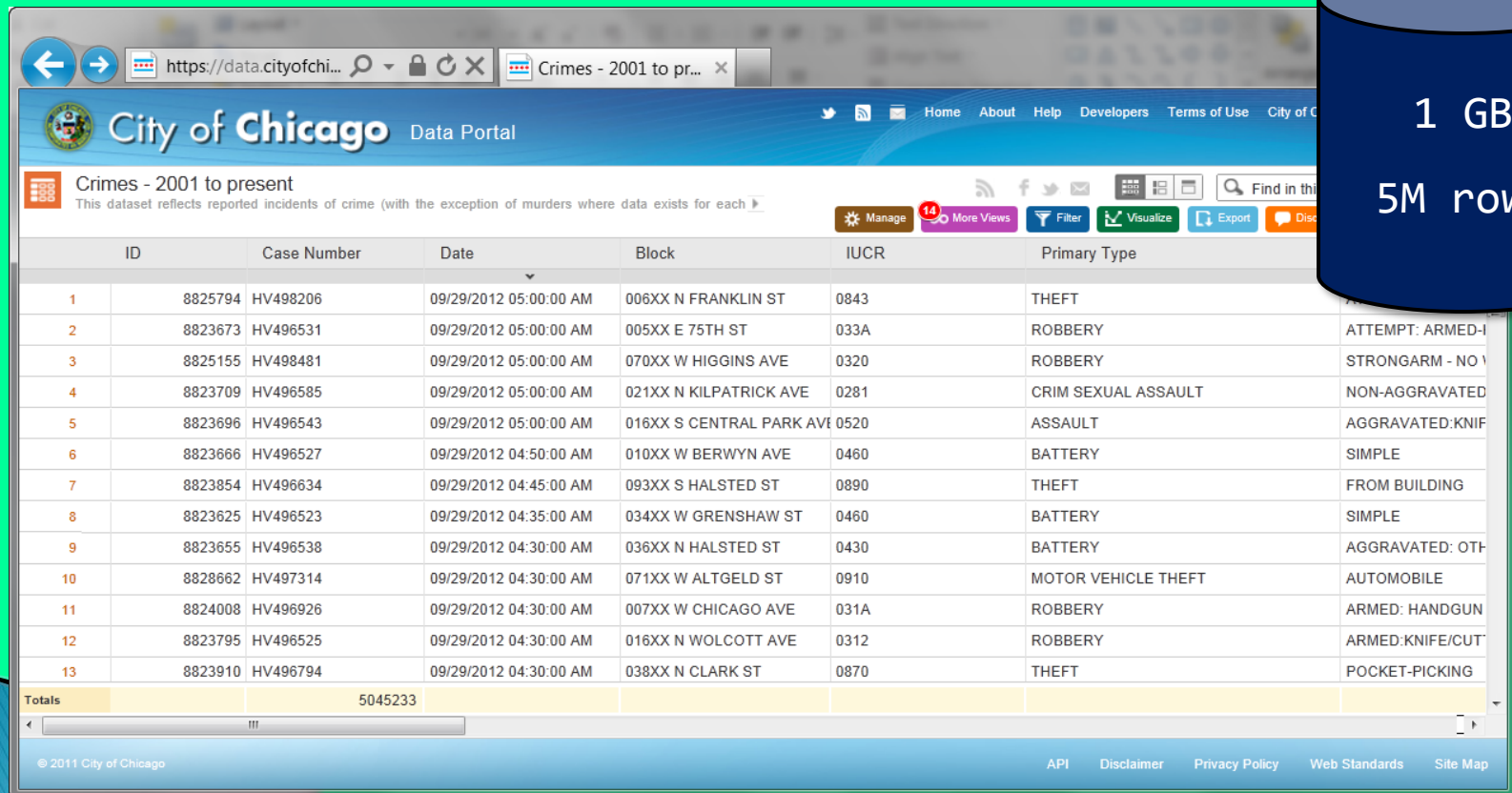
Compare and contrast

Approach	Pros	Cons	Target	Scalable?
PowerPivot	No programming, built-in UI and visualization	Lack of scalability	GBs	Limited by RAM
LINQ	Flexibility of analysis	Programming	GBs, few TBs	Limited by local resources
Hadoop	Scalability, ease of programming	Must fit into Map-Reduce framework; not necessarily fast	GBs, TBs, PBs	Yes! (via cluster or cloud)

Data set for talk

▶ We'll be working with Chicago crime data...

- <https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2>
- <http://www.cityofchicago.org/city/en/narr/foia/CityData.html>



City of Chicago Data Portal

Crimes - 2001 to present

This dataset reflects reported incidents of crime (with the exception of murders where data exists for each)

ID	Case Number	Date	Block	IUCR	Primary Type		
1	8825794	HV498206	09/29/2012 05:00:00 AM	006XX N FRANKLIN ST	0843	THEFT	
2	8823673	HV496531	09/29/2012 05:00:00 AM	005XX E 75TH ST	033A	ROBBERY	ATTEMPT: ARMED-I
3	8825155	HV498481	09/29/2012 05:00:00 AM	070XX W HIGGINS AVE	0320	ROBBERY	STRONGARM - NO V
4	8823709	HV496585	09/29/2012 05:00:00 AM	021XX N KILPATRICK AVE	0281	CRIM SEXUAL ASSAULT	NON-AGGRAVATED
5	8823696	HV496543	09/29/2012 05:00:00 AM	016XX S CENTRAL PARK AV	0520	ASSAULT	AGGRAVATED: KNIF
6	8823666	HV496527	09/29/2012 04:50:00 AM	010XX W BERWYN AVE	0460	BATTERY	SIMPLE
7	8823854	HV496634	09/29/2012 04:45:00 AM	093XX S HALSTED ST	0890	THEFT	FROM BUILDING
8	8823625	HV496523	09/29/2012 04:35:00 AM	034XX W GRENSHAW ST	0460	BATTERY	SIMPLE
9	8823655	HV496538	09/29/2012 04:30:00 AM	036XX N HALSTED ST	0430	BATTERY	AGGRAVATED: OTH
10	8828662	HV497314	09/29/2012 04:30:00 AM	071XX W ALTGELD ST	0910	MOTOR VEHICLE THEFT	AUTOMOBILE
11	8824008	HV496926	09/29/2012 04:30:00 AM	007XX W CHICAGO AVE	031A	ROBBERY	ARMED: HANDGUN
12	8823795	HV496525	09/29/2012 04:30:00 AM	016XX N WOLCOTT AVE	0312	ROBBERY	ARMED:KNIFE/CUT
13	8823910	HV496794	09/29/2012 04:30:00 AM	038XX N CLARK ST	0870	THEFT	POCKET-PICKING
Totals		5045233					

© 2011 City of Chicago

API Disclaimer Privacy Policy Web Standards Site Map

1 GB
5M rows

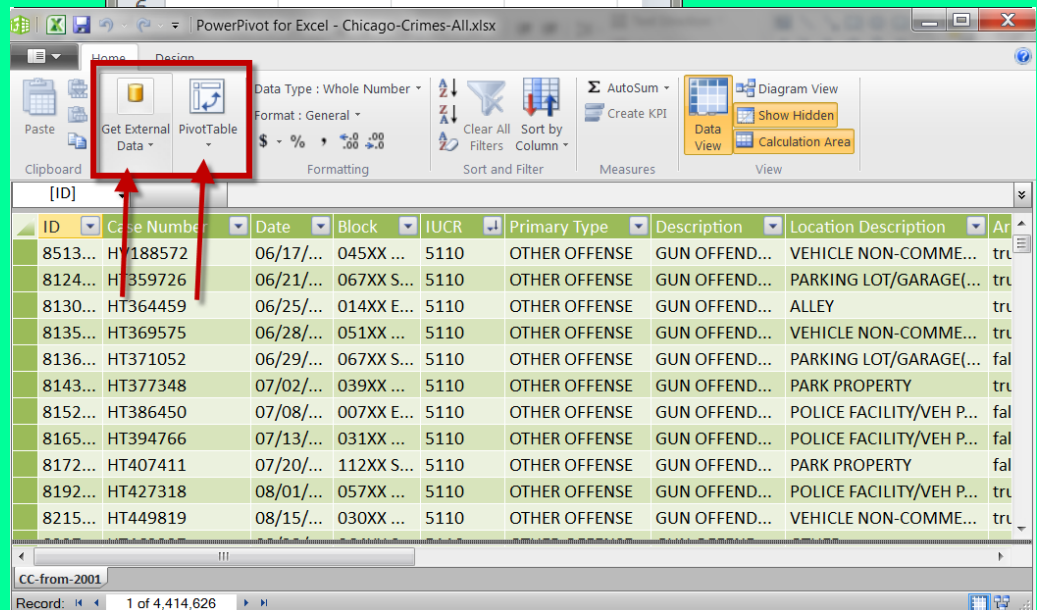
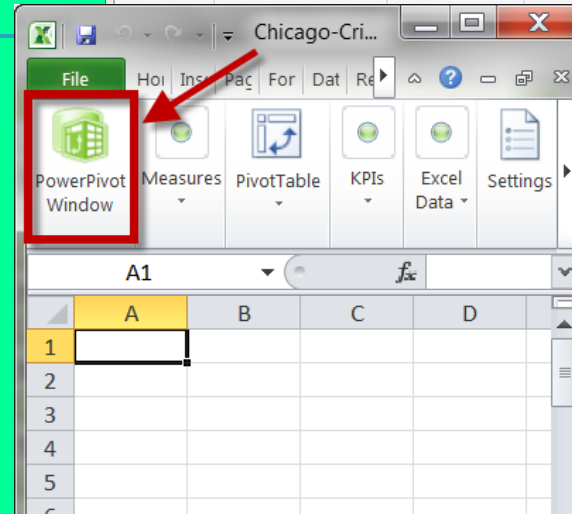
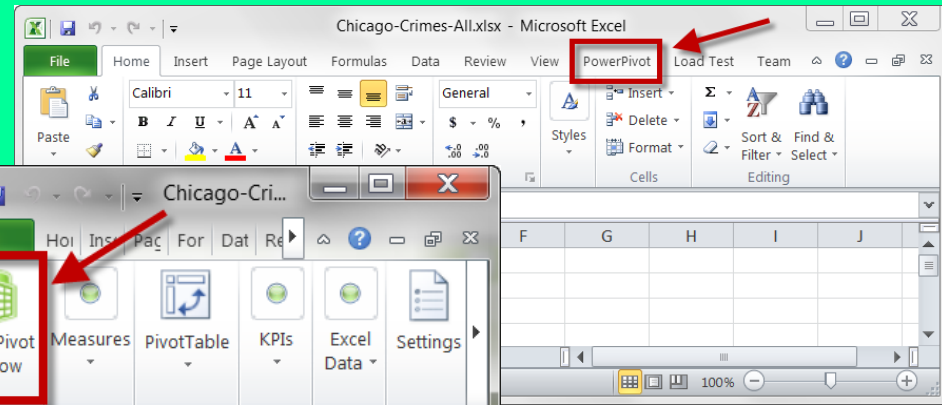
(1) PowerPivot

- ▶ Freely-available plugin for Excel 2010
 - <http://www.powerpivot.com/>
- ▶ Turns Excel into an in-memory database
 - *More precisely, turns spreadsheet into an OLAP cube*
- ▶ Note:
 - *If you have 32-bit Excel, install 32-bit PowerPivot*
 - *If you have 64-bit Excel, install 64-bit PowerPivot*
 - *GBs of data will require 64-bit*
 - *[How to tell what version of Excel you have? File menu, help...]*

Demo

▶ PowerPivot...

- *Install*
- *PowerPivot menu*
- *PowerPivot Window*
- *Get Data...*
- *PivotTable...*



(2) LINQ

- ▶ LINQ == *Language Integrated Query*
- ▶ Traditional programming + SQL
 - <http://msdn.microsoft.com/en-us/library/bb397926.aspx>
 - <http://code.msdn.microsoft.com/101-LINQ-Samples-3fb9811b>
- ▶ Included with .NET, which is freely-available:
 - Windows: *Microsoft .NET Framework SDK*
 - Linux, Mac, Windows: *Mono project*

Demo

▶ LINQ...

// Using LinqToCsv package:

```
IEnumerable<CrimeReport> crimes =  
    csv.Read<CrimeReport>("CC-from-2001.txt", new CsvFileDescription {  
        SeparatorChar = ',', FirstLineHasColumnNames = true});
```

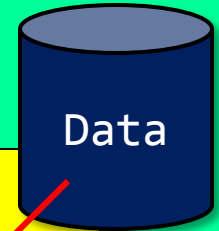
```
var query =
```

```
    from crime in crimes
```

```
    group crime by crime.IUCR into g
```

```
    orderby g.Count() descending
```

```
    select new { IUCR = g.Key, Count = g.Count() };
```




[<code1,count>, <code2,count>, ...]

0486 366903
0820 308074
.
.
.

More efficient version (10x)

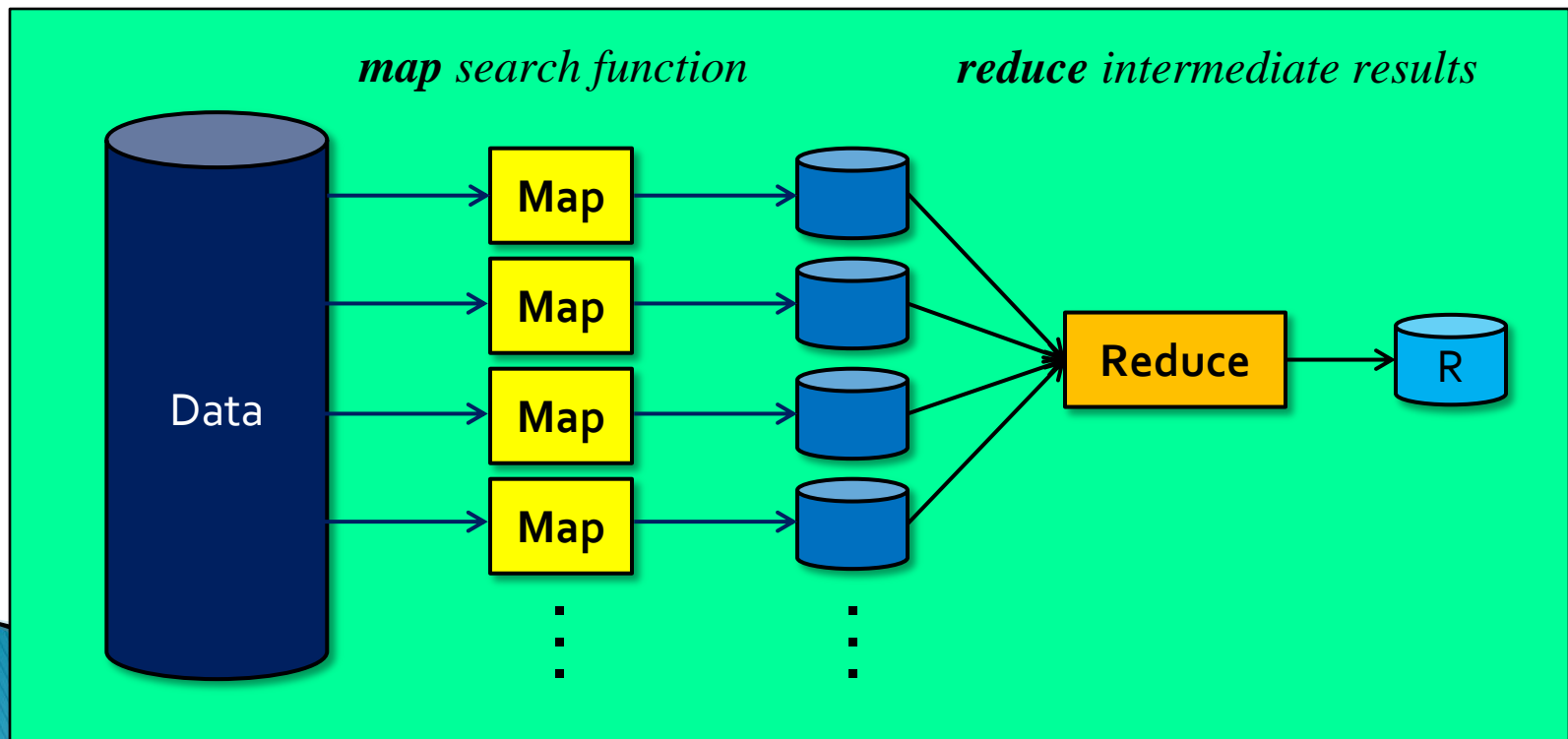
- ▶ Parse data inline...

```
var query =  
    File.ReadLines("CC-from-2001.txt").Skip(1).AsParallel().  
    Select(line =>  
    {  
        string[] values = line.Split(new char[] { ',' });  
        return (values.Length > 4) ? values[4] : "?????";  
    }  
    ).  
    GroupBy(iucr => { return iucr; }).  
    OrderBy(g => { return -g.Count(); }).  
    Select(g => new { IUCR = g.Key, Count = g.Count() });
```

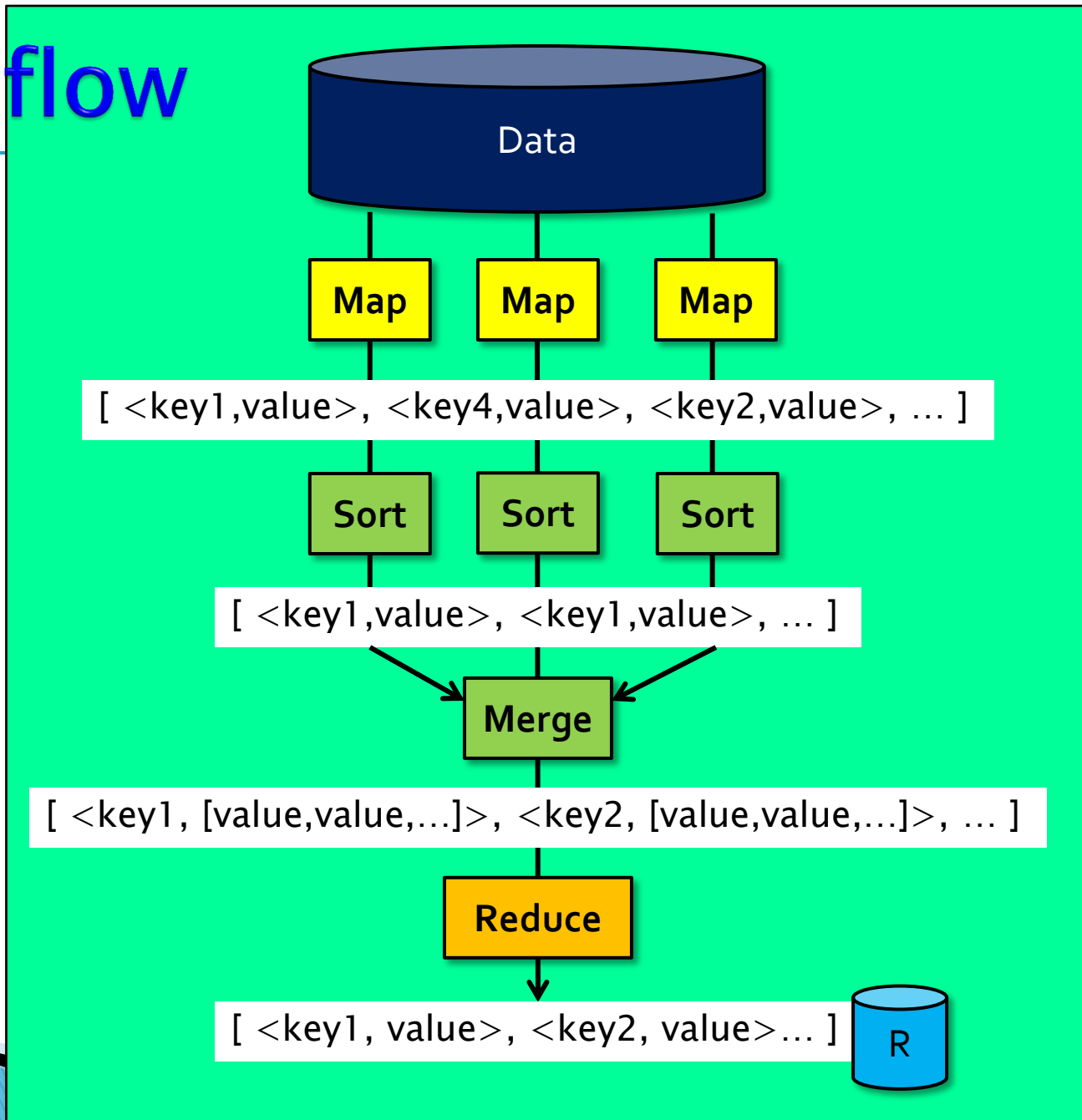


(3) Hadoop

- ▶ Freely-available framework for big data
 - <http://hadoop.apache.org/>
- ▶ Based on concept of Map-Reduce:



Workflow



Demo

▶ Hadoop on Azure...

// Javascript version:

```
var map = function (key, value, context)
{
    var values = value.split(",");
    context.write(values[4], 1);
};

var reduce = function (key, values, context)
{
    var sum = 0;
    while (values.hasNext())
    {
        sum += parseInt(values.next());
    }
    context.write(key, sum);
};
```

Hadoop++

▶ Rich ecosystem around Hadoop

- Pig
- Hive
- HBASE
- ...

// interactive PIG with explicit Map-Reduce functions:

```
pig.from("CC-from-2001.txt").  
  mapReduce("IUCR-Count.js", "IUCR, Count:long").  
  orderBy("Count DESC").  
  take(10).  
  to("output-from-2001")
```

// interactive PIG without explicit Map-Reduce:

```
schema = "ID,Case Number,Date,Block,IUCR,..."  
pig.from("CC-from-2001.txt", schema).  
  groupBy("IUCR").  
  select("group, SUM($1.Count)").  
  orderBy("Count DESC").  
  take(10).  
  to("output-from-2001")
```

Hadoop on Azure

- ▶ Microsoft is offering free access to Hadoop
 - Request invitation @ <http://www.hadooponazure.com/>

- ▶ Hadoop connector for Excel
 - *Process data using Hadoop, analyze/visualize using Excel*

That's it!

Thank you for attending



- ▶ Presenter: Joe Hummel

- *Email:* joe@joehummel.net
- *Materials:* <http://www.joehummel.net/downloads.html>

- ▶ Upcoming products of interest...

- **PowerView:**
 - *Plugin for Excel 2013*
- **Hadoop on Azure**
- **Hadoop on Windows**