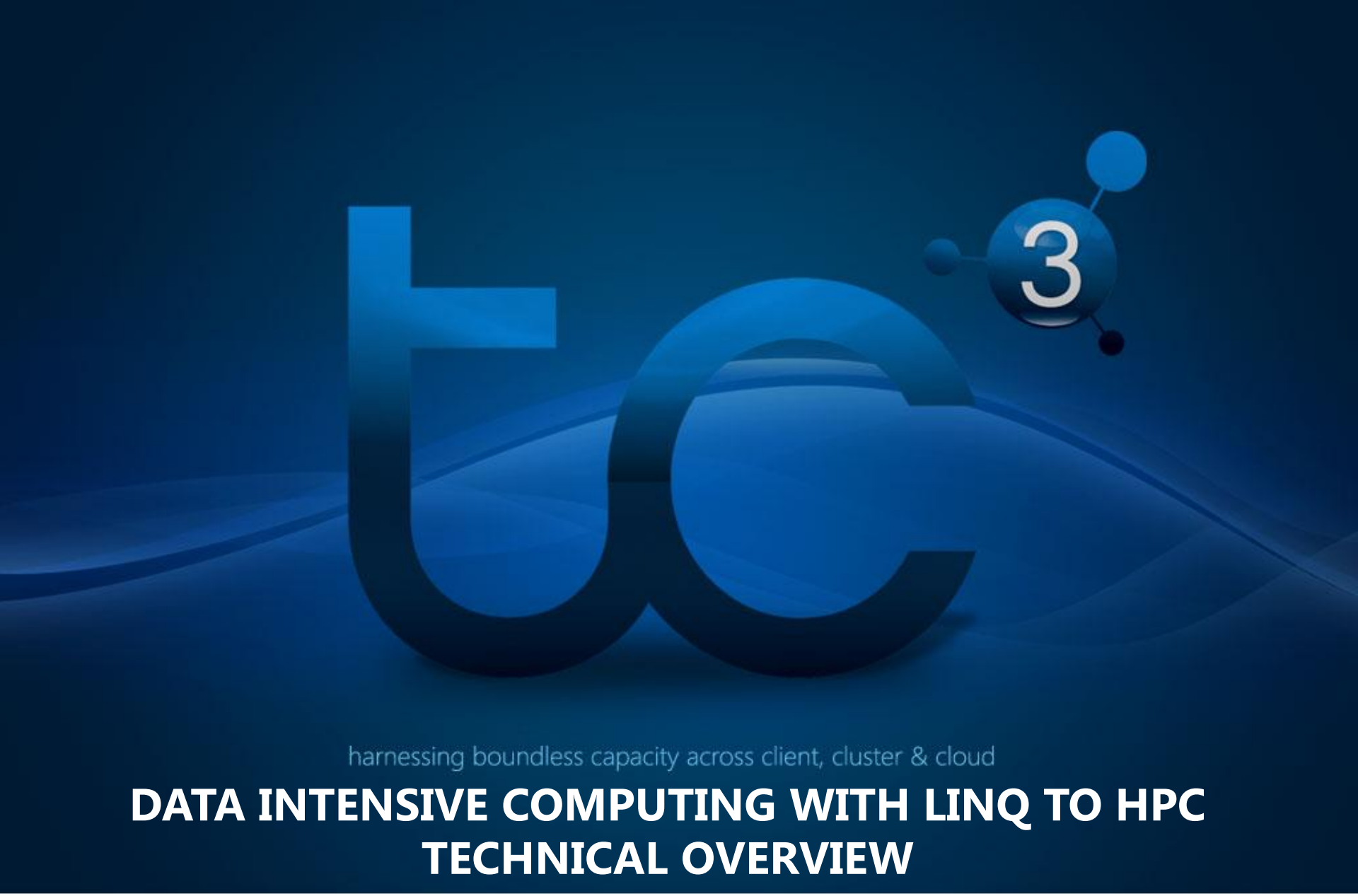


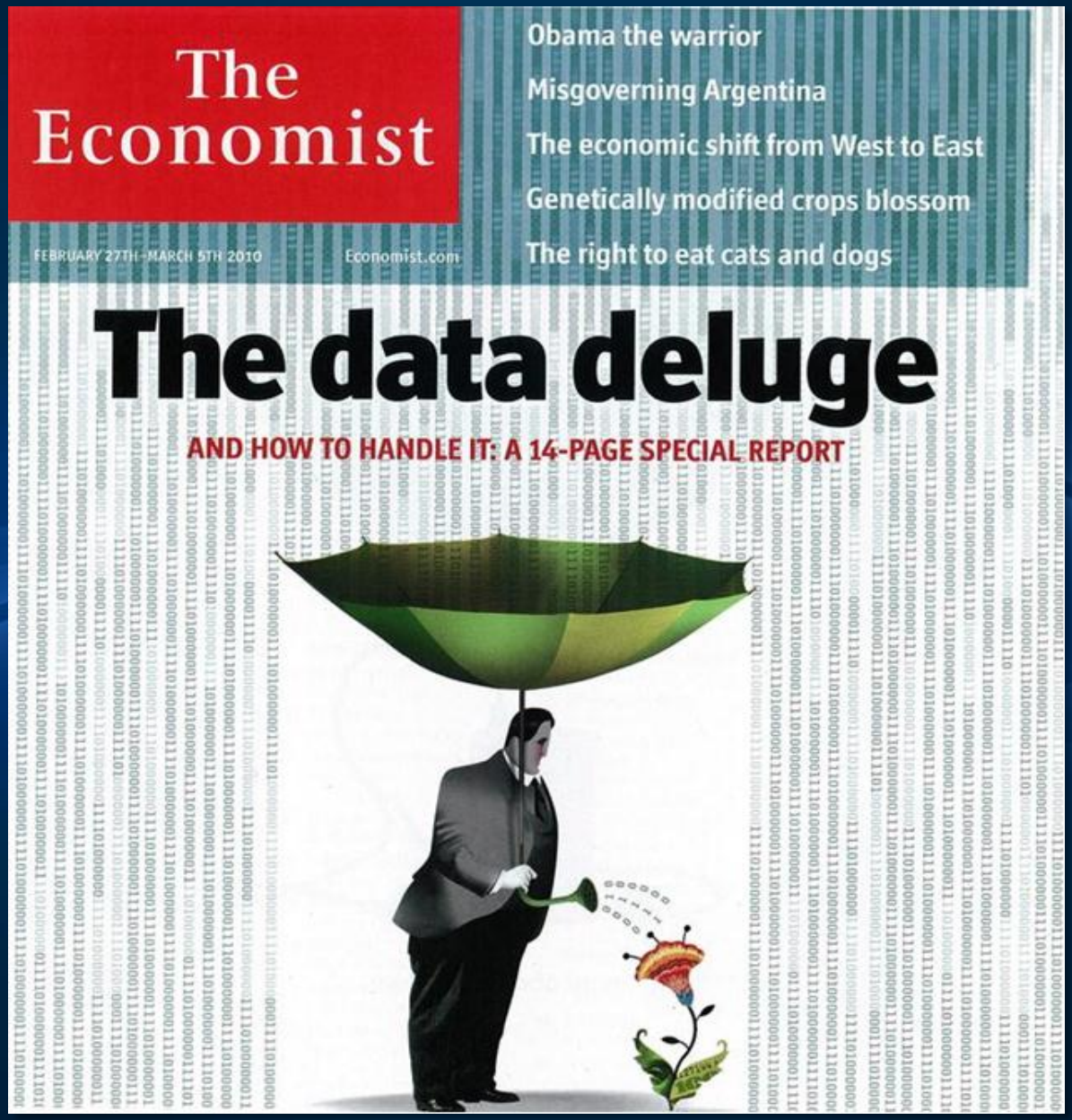
DATA INTENSIVE COMPUTING WITH LINQ TO HPC TECHNICAL OVERVIEW

Ade Miller, Principal Program Manager, LINQ to HPC

Microsoft Visual Studio 2010 | Windows HPC Server 2008 R2 | Microsoft Excel 2010 | Windows Azure

Agenda

- Introduction to LINQ to HPC
- Using LINQ to HPC
- Systems Management
- Integrating with Other Data Technologies


The Economist
Obama the warrior
Misgoverning Argentina
The economic shift from West to East
Genetically modified crops blossom
The right to eat cats and dogs
FEBRUARY 27TH–MARCH 5TH 2010
Economist.com
The data deluge
AND HOW TO HANDLE IT: A 14-PAGE SPECIAL REPORT

The Data Spectrum

- One extreme is analyzing raw, unstructured data. The analyst does not know exactly what the data contains, nor what cube would be justified. The analyst needs to do ad-hoc analyses that may never be run again.

LINQ to HPC

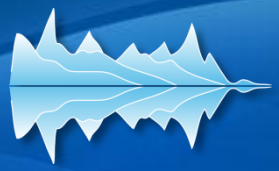
- Another extreme is analytics targeting a traditional data warehouse. The analyst knows the cube he or she wants to build, and the analyst knows the data sources.

Parallel Data Warehouse

What kind of Data?

Large Data Volume

- 100s of TBs to 10s of PBs



Non-Traditional data Types

- Unstructured & Semi structured
- Weak relational schema
- Text, Images, Videos, Logs

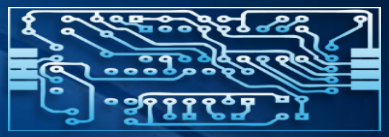
New Data Sources

- Sensors & Devices
- Traditional applications
- Web Servers
- Public data

Q+i

New Questions & New Insights

- How popular is my product?
- What is the best ad to serve?
- Is this a fraudulent transaction?



New Technologies

- Distributed Parallel Processing Frameworks
- Easy to Scale on commodity hardware
- MapReduce-style programming models

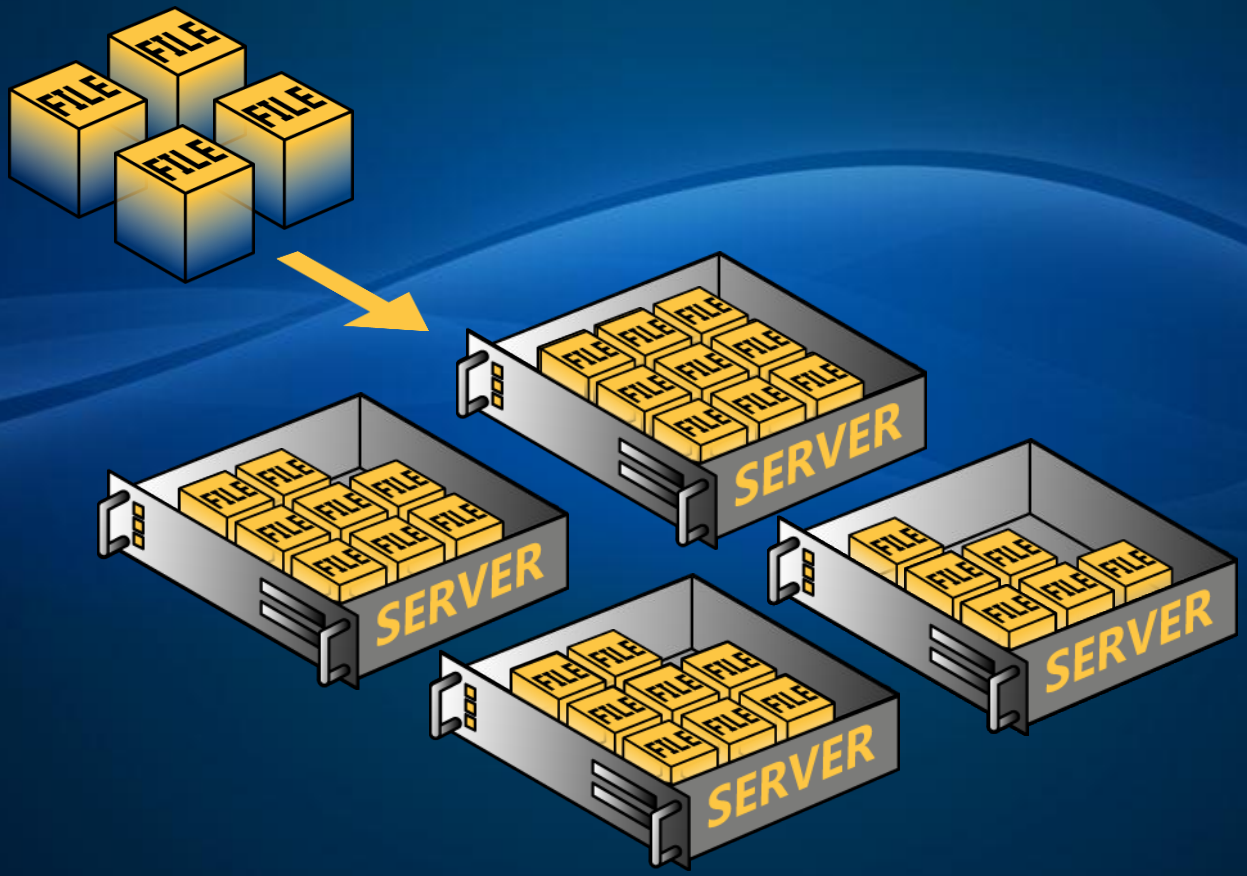
Overview

MOVING THE DATA TO THE COMPUTE

So how does it work?

FIRST, STORE THE DATA

Data Intensive Computing with HPC Server

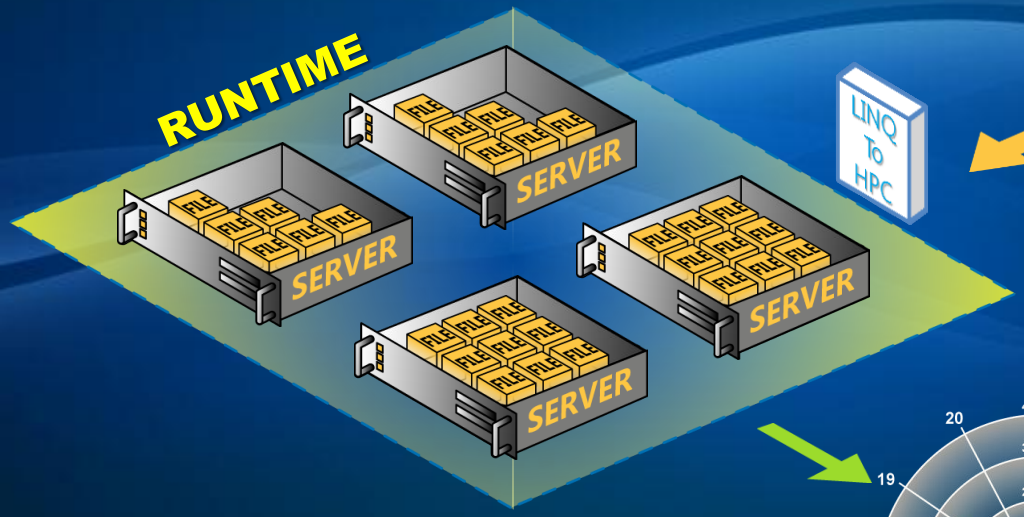


Windows HPC Server 2008 R2 Service Pack 2

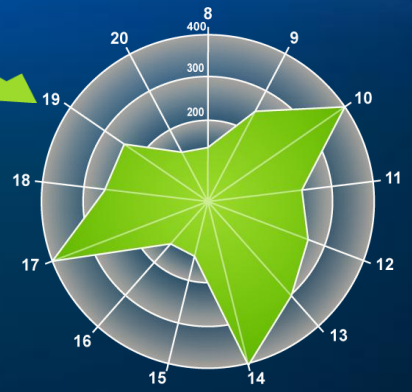
So how does it work?

SECOND, TAKE THE PROCESSING TO THE DATA

Data Intensive Computing with HPC Server




```
var logentries =
        from line in logs
        where !line.StartsWith("#")
        select new LogEntry(line);
var user =
        from access in logentries
        where access.user.EndsWith(@"\Ade")
        select access;
var accesses =
        from access in user
        group access by access.page into pages
        select new UserPageCount("Ade",
                                 pages.Key,
pages.Count());
var htmAccesses =
        from access in accesses
        where access.page.EndsWith(".htm")
        orderby access.count descending
        select access;
```



Microsoft® Visual Studio®

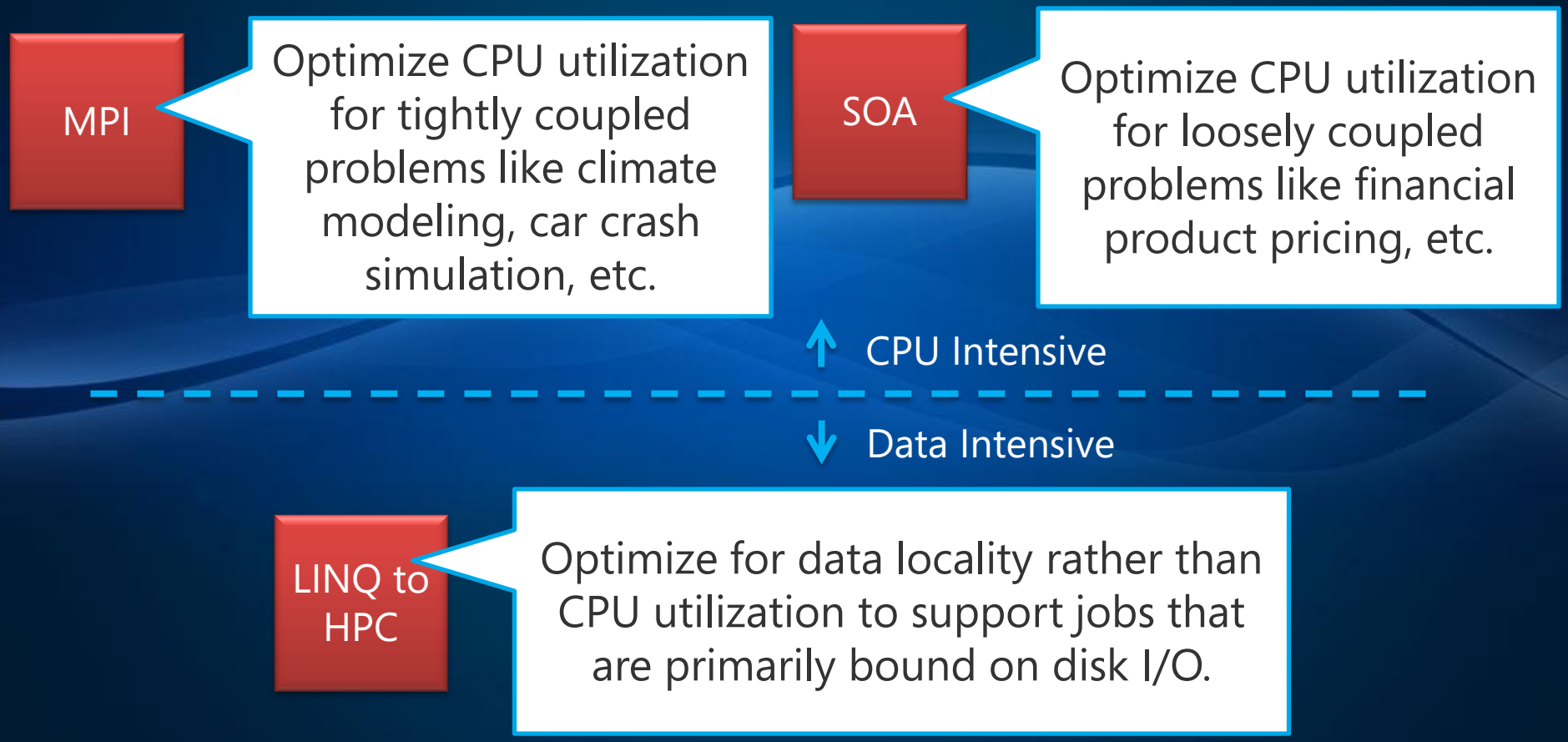
Data Intensive Computing with HPC Server 2008 R2

INTRODUCTION TO LINQ TO HPC

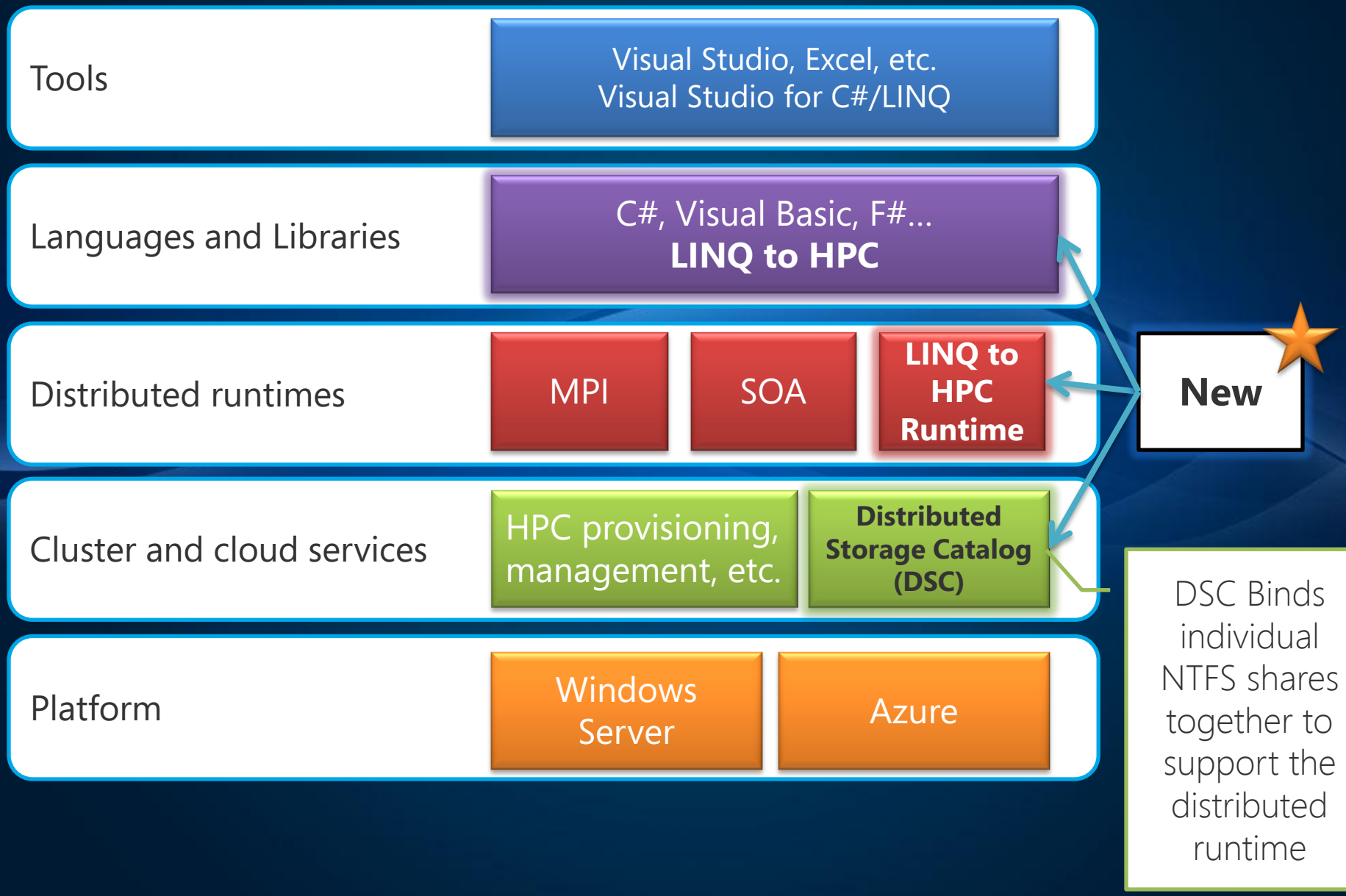
History of LINQ to HPC

- Developed by Microsoft Research as “Dryad”
- Same technologies used internally within Microsoft
  - Powered Microsoft Search’s analytics since August 2006
  - Scaled to 10K servers in a single cluster
- Build on existing technologies
  - SQL Server
  - The NTFS file system
- Now being delivered as part of HPC Pack

Solving a new class of problems



Building blocks



Example: find web pages from many log files

LINQ query transformed into computation graph

```
var logentries =
      from line in logs
      where !line.StartsWith("#")
      select new LogEntry(line);
var user =
      from access in logentries
      where access.user.EndsWith(@"\Ade")
      select access;
var accesses =
      from access in user
      group access by access.page into pages
      select new UserPageCount("Ade", pages.Key,
            pages.Count());
var htmAccesses =
      from access in accesses
      where access.page.EndsWith(".htm")
      orderby access.count descending
      select access;
```

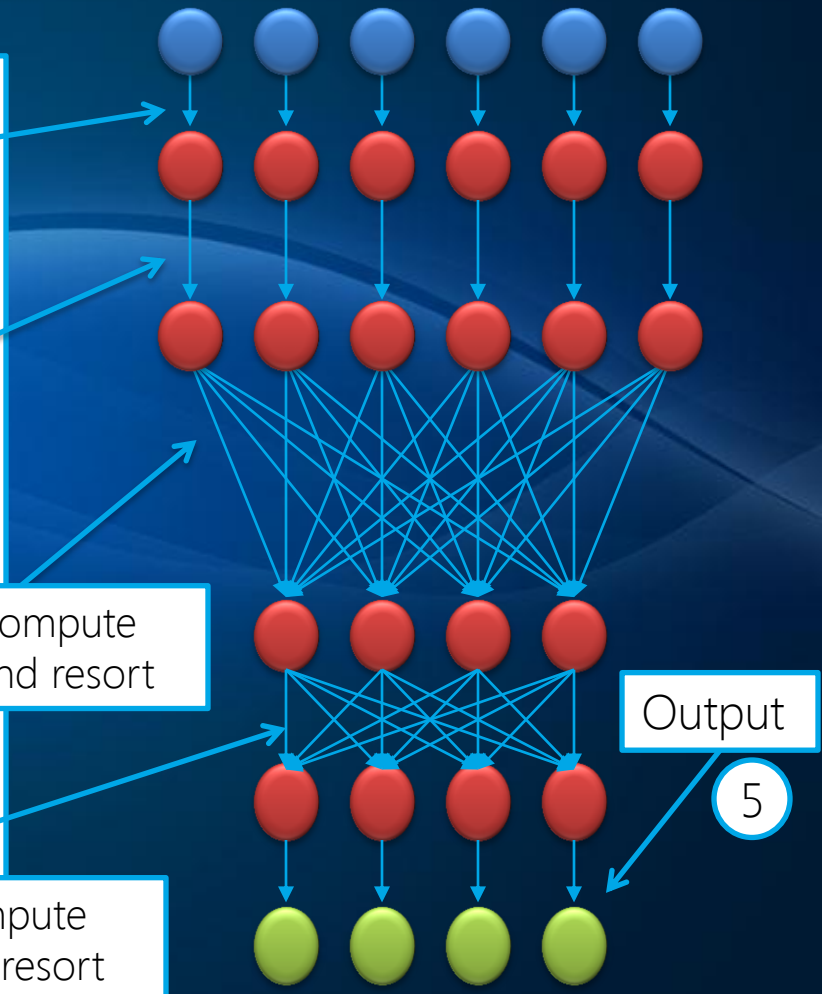
1 Input

Compute 2

3 Compute and resort

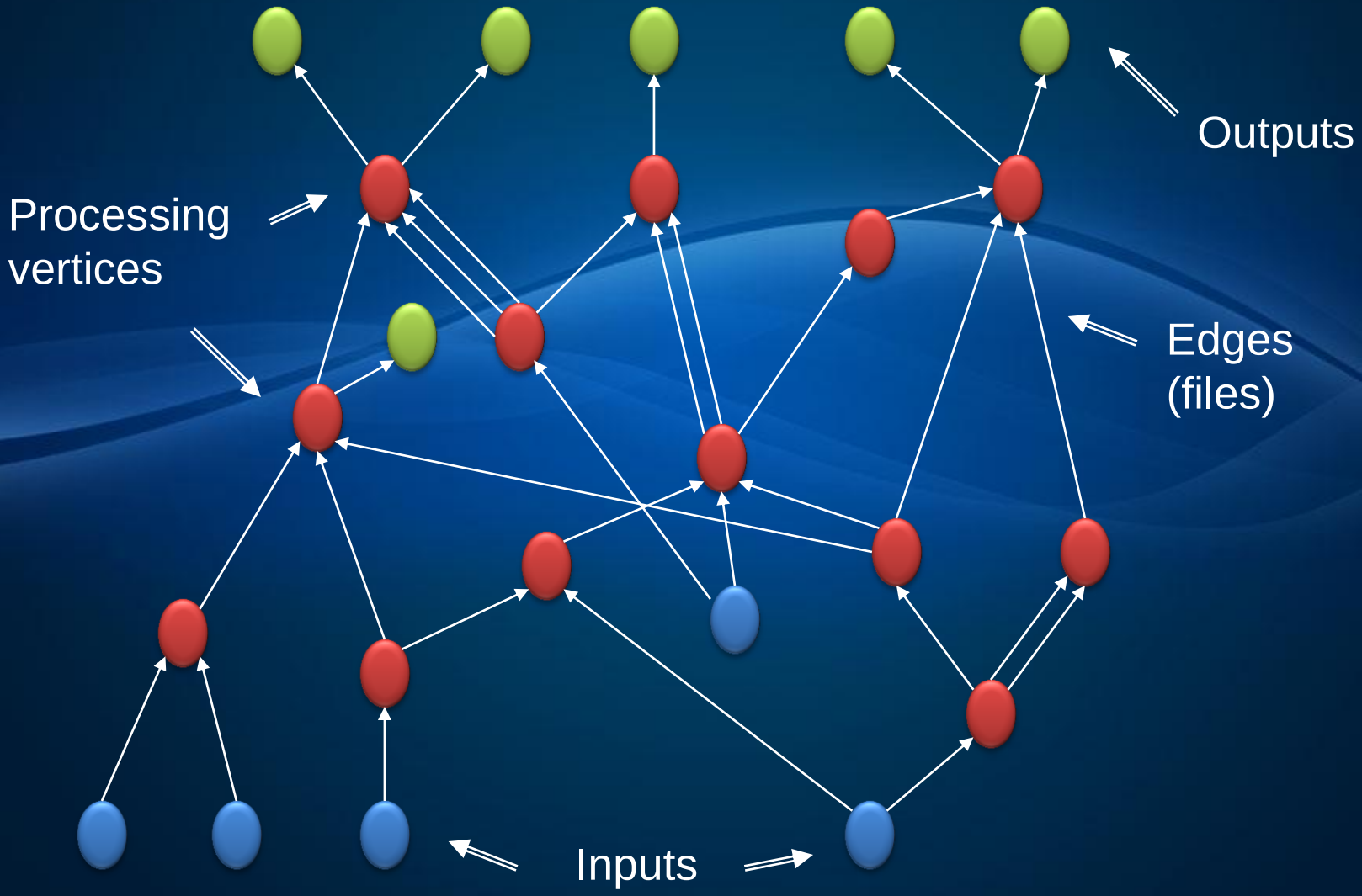
4 Compute and resort

Output 5



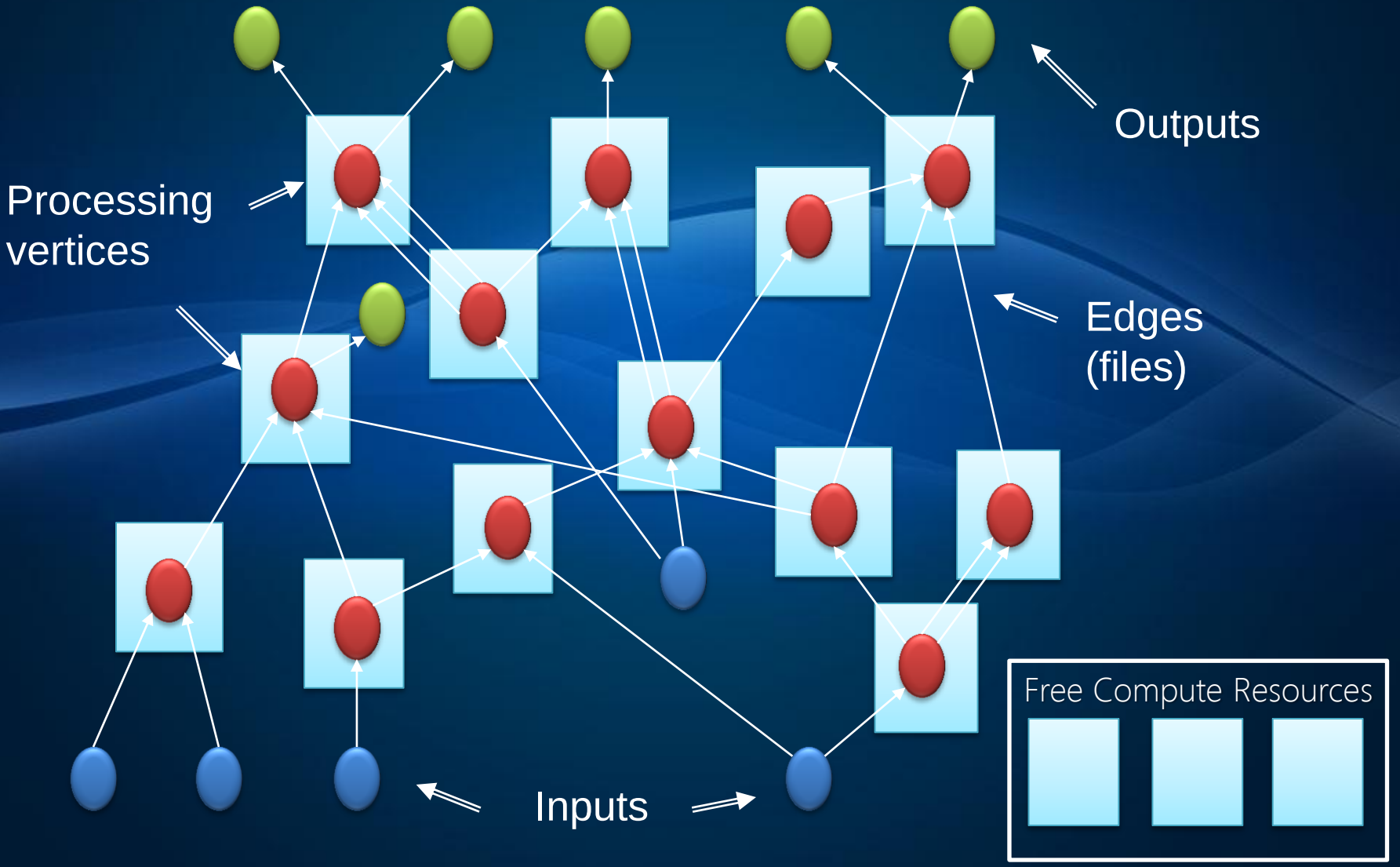
LINQ to HPC

DIRECTED ACYCLIC GRAPH (DAG) OF VERTICES



Execute DAGs

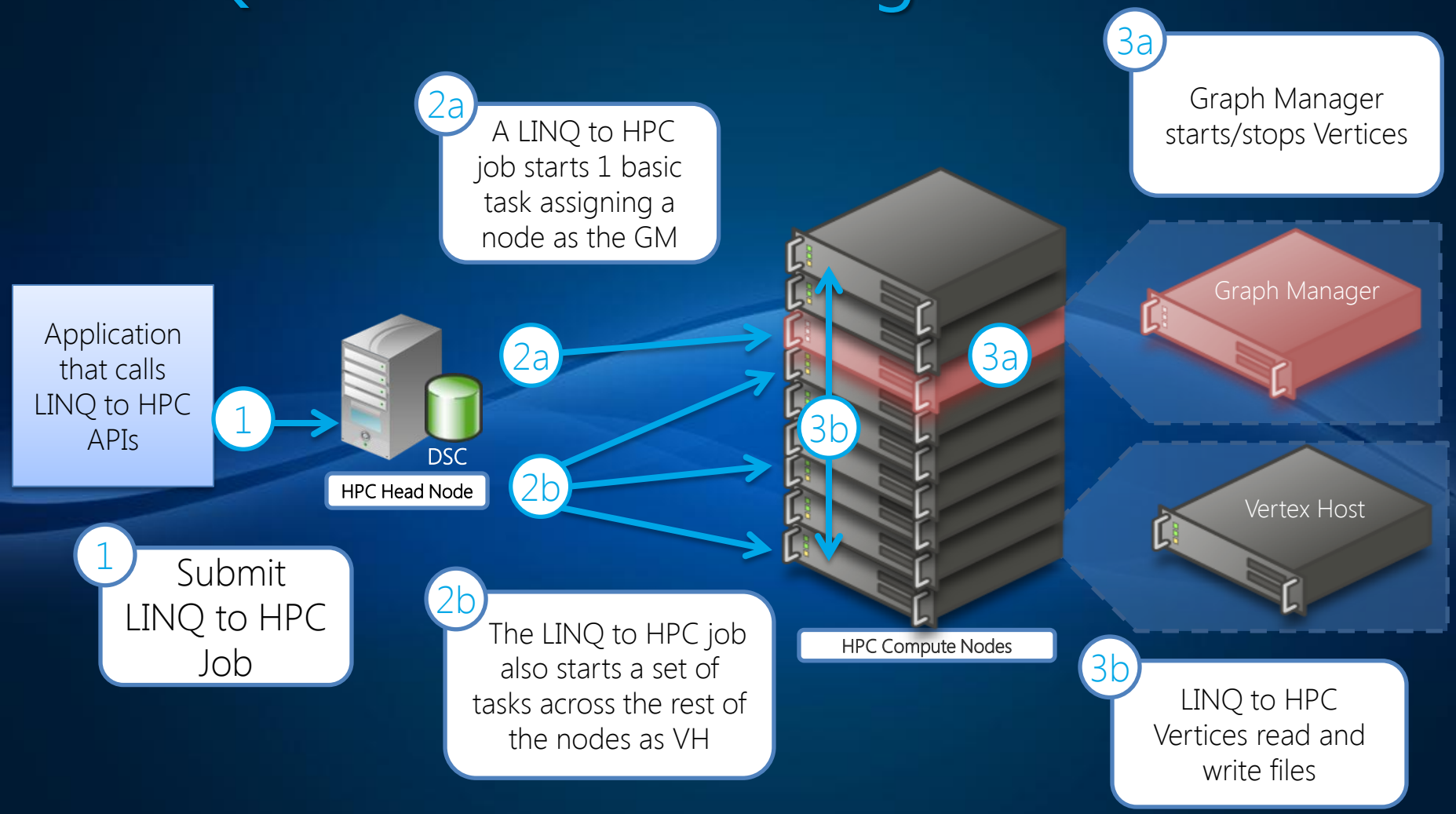
MAPPING VERTICES TO DISTRIBUTED VERTEX HOSTS



Data Intensive Computing with HPC Server 2008 R2

LINQ TO HPC WALKTHROUGH

LINQ to HPC Walkthrough



HPC + LINQ to HPC Mechanics

2a
A LINQ to HPC job starts 1 basic task assigning a node as the GM

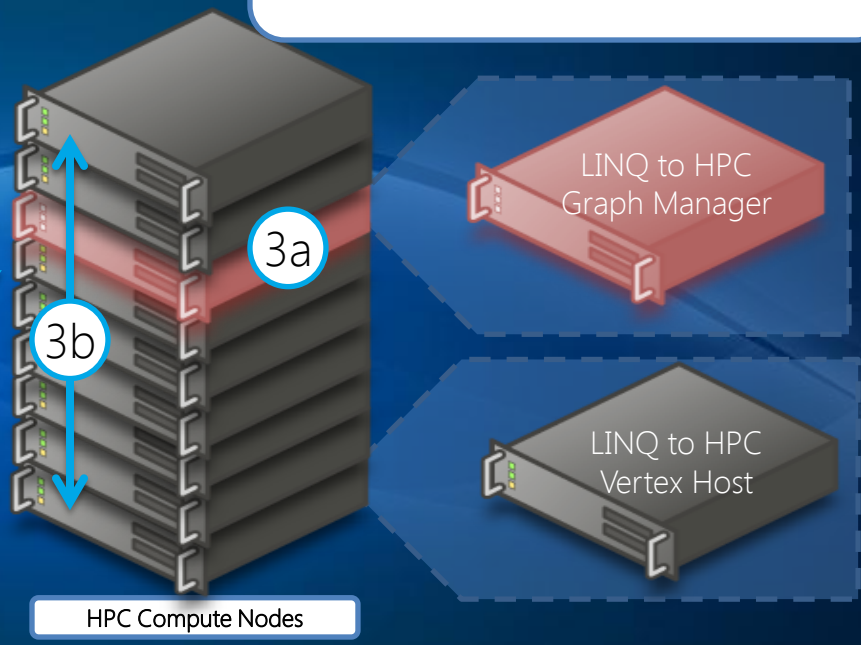
3a
GM reads XML description of graph from share, calls DSC to locate files referenced in XML

Application that calls LINQ to HPC APIs

1

2a

2b



1
Publish to share:
1. binaries for LINQ to HPC job
2. XML description of LINQ to HPC graph

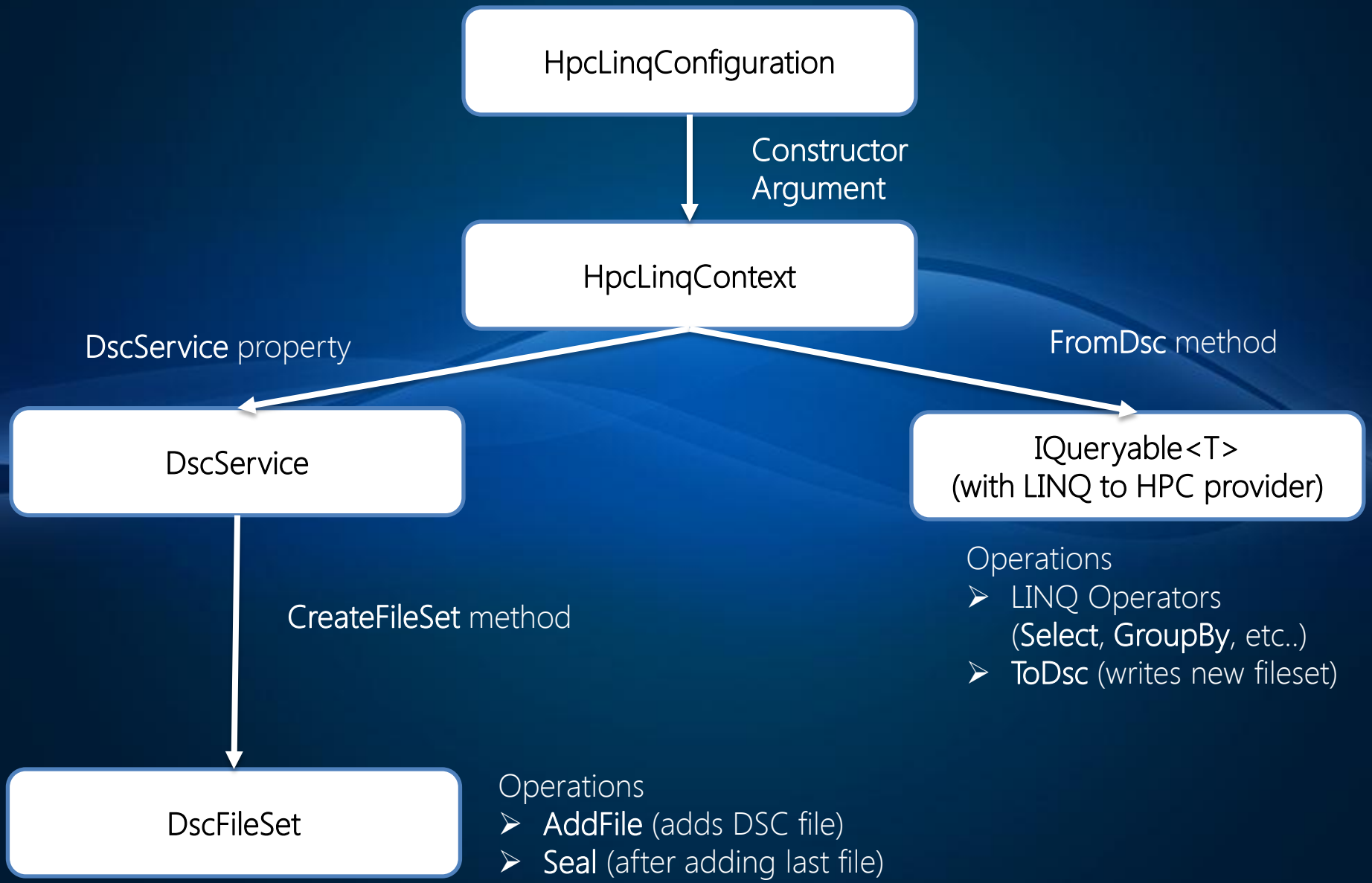
2b
The LINQ to HPC job also starts a set of tasks across the rest of the nodes as VH

3b
VH loads binaries for this LINQ to HPC job from share, executes them according to commands from GM

Data Intensive Computing with HPC Server 2008 R2

CODING WITH LINQ TO HPC

LINQ to HPC Object Model



Getting Started

HELLO WORLD!

```
using System;
using System.Linq;
using Microsoft.Hpc.Linq;

namespace MyProgram {
  class Program {
    static void Main(string[] args) {

      var config = new HpcLinqConfiguration("HEADNODE");
      var context = new HpcLinqContext(config);

      var max = context.FromDsc<LineRecord>("MyTextData")
        .Select(r => r.Line.Length)
        .Max();

      Console.WriteLine("The max line length is " + max);
    }
  }
}
```

Detailed Code Examples

EXAMPLES: LINQ, LINQ TO HPC (LOCAL), LINQ TO HPC (DSC)

Examples
SORT

```
var config = new HpcLinqConfiguration("HEADNODE");
HpcLinqContext context = new HpcLinqContext(config));


context.FromDsc<LineRecord>("input")
   .OrderBy(r => r.Line, new MyComparer(10))
   .ToDsc("sorted output")
   .SubmitAndWait(context);
```

Examples
WORD COUNT

```
var config = new HpcLinqConfiguration("HEADNODE");
HpcLinqContext context = new HpcLinqContext(config));


IQueryable<Pair> results =
  context.FromDsc<LineRecord>("input")
    .SelectMany(line => line.Line.Split(new[] {' ', '\t'}))
    .GroupBy(word => word)
    .Select(word => new Pair(word.Key, word.Count()))

    .OrderByDescending(pair => pair.Count)
    .Take(200);
```

Examples

WORD COUNT (WITH MAPREDUCE)

```
Expression<Func<LineRecord, IEnumerable<string>>> mapper =
  (line) => line.Line.Split(new[] { ' ', '\t' });

Expression<Func<string, string>> selector = word => word;

Expression<Func<string, IEnumerable<string>, Pair>> reducer =
  (key, words) => new Pair(key, words.Count());

IQueryable<Pair> results =
  context.FromDsc<LineRecord>("input")
    .MapReduce(mapper, selector, reducer)

    .OrderByDescending(pair => pair.Count)
    .Take(200);
```

ADDITIONAL TOOLS

- Profiling Tools
  - View Query Plan
  - Profile query and stage timings
- Distributed Storage Catalog Explorer
- Command Line Tools
  - CMD & PowerShell

Administration

DEPLOYMENT AND SYSTEMS MANAGEMENT

Managing Data and HPC Server

- HPC Server administration basics:
  - Managing the job queue
  - How to identify the user that submitted jobs
  - Canceling a runaway job
- Data Storage Catalog specific tasks:
  - Monitor disk usage tracked by DSC on each node
  - View how the DSC file set maps to NTFS across nodes
  - Identify the nodes where files are replicated
  - Add and remove data from the cluster

Integration

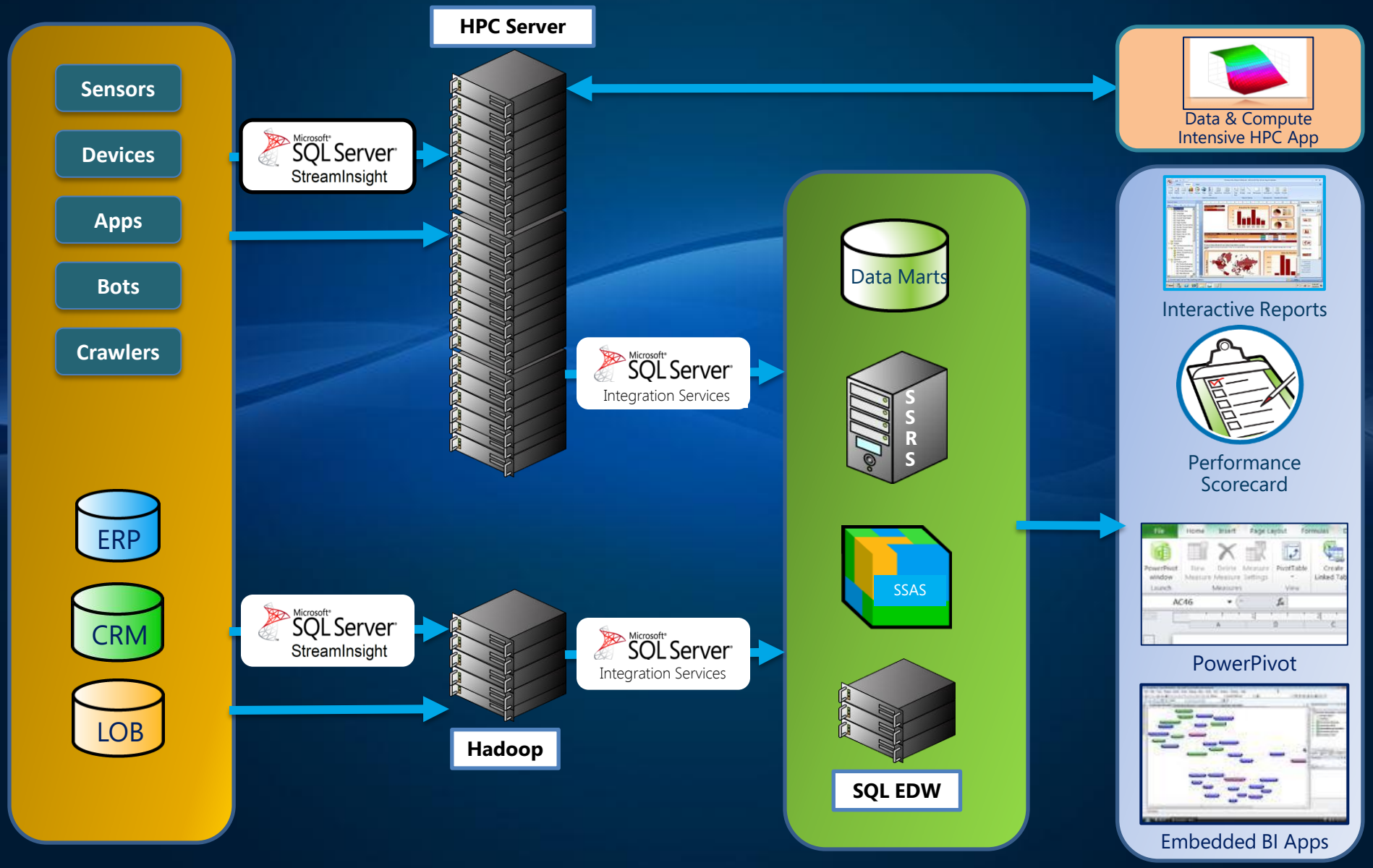
COMBINING HPC SERVER AND YOUR OTHER DATA ASSETS

Microsoft has great data platform assets

- Microsoft already has great data platform assets

  PowerPivot, SQL Server Integration Services (SSIS), Parallel Data Warehouse (PDW), …
- LINQ to HPC focuses on **raw unstructured data analytics** enables new solutions that incorporate multiple assets
  - E.g., analyze raw unstructured data using LINQ to HPC then pipe it to SSIS and apply rest of BI stack

End-to-End Data Intensive Computing



For more Information

- Download HPC Server 2008 R2 Evaluation Copy Today – microsoft.com/hpc
- Download Service Pack 2
- Download LINQ to HPC Beta 2 - connect.microsoft.com
- Try HPC Server Hands-on Labs – microsoft.com/hpc -> Technical Resources

Microsoft®