TPC-H Analtycs' Scenarios and Performances on Hadoop Data Clouds

Rim Moussa
rim.moussa@esti.rnu.tn
LATICE –Univ. of Tunis
Tunisia

24th, April 2012

4th International. Conference on Networked Digital Technologies NDT'2012, Dubai, UAE.

OUTLINE

1. Business Intelligence
2. Motivation
   - data managment issues, NoSQL, clouds
   - OLAP in the cloud
3. Implementation of OLAP in the cloud
   - TPC-H Benchmark
   - Analytics Scenarios
   - Performance Measurements
4. Related Work
5. Conclusion
6. Future Work

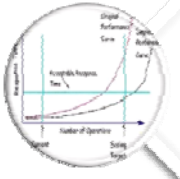
BUSINESS INTELLIGENCE

- Business intelligence aims to support better business decision-making.
- Common functions of business intelligence technologies are
  - **On-Line Analytical Processing,**
  - data mining, process mining,
  - Business performance management
  - Text mining and predictive analytics, …
- Market share
  - Gartner Research Reports BI Market Revenue Hit **$12.2 Billion** in 2011

MOTIVATION

Decision Support Systems

- Incessant Data & complex workload
- Complex DB schema



Scalability Issues

- Ideally, Linear Speed up & Linear Scale up
- DBMS do not scale linearly
- OLAP Technologies do not scale

Hardware

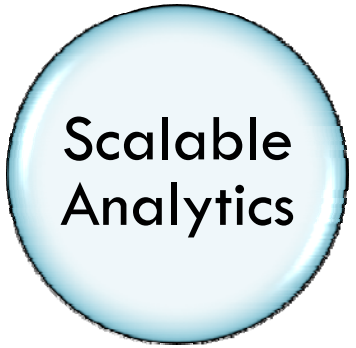
- I/O Bottleneck → I/O-bound data storage systems
- Gilder law: Thrice bandwidth every 3 years
- Moore Law: Twice computing and storage capacities every 18 months. Obsolete by 2017
- Vertical scaling cost >> Horizontal scaling cost

NoSQL

- Big Challenges related to **velocity**
  - How fast huge volumes of data can be processed?
- NoSQL solutions
  - Adopted by Google, Facebook, Amazon, …
  - Dynamic horizontal scale-up
    - Nodes are added without bringing the cluster down
    - Shared-nothing architecture
    - Independent computing& storage nodes interconnected via a high speed network
  - Distributed programming framework: MapReduce (Google)

CLOUD COMPUTING

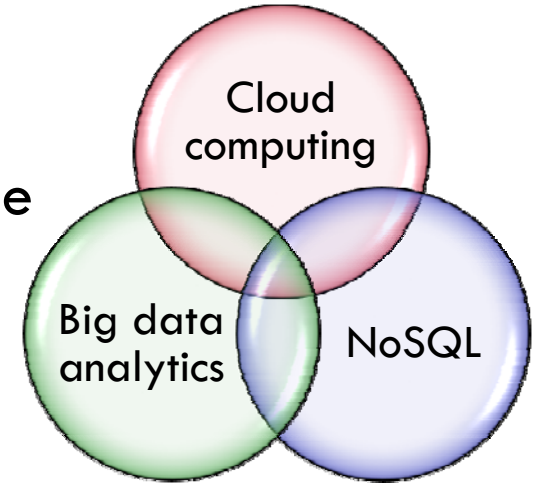
- *Cloud computing is a style of computing where scalable and elastic IT-enabled capabilities are provided "as a service" to external customers using Internet technologies.*
  - Broad network access
  - Resource pooling (virtualization)
  - Self-provisioning
  - Rapid elasticity
  - Measured service



- Market share
  - Forrester Research expects the global cloud computing market to reach $241 billion in 2020. In particular, SaaS market growing to $92.8 billion by 2016.
  - Gartner group expects the cloud computing market will reach $US150.1 billion, with a compound annual rate of 26.5%, in 2013.

OLAP IN THE CLOUD

- OLAP constraints
  - Big data analytics' obstacles
  - Current systems & technologies do not scale
- Key benefits of Cloud Computing
  - Performance
    - Much faster data analysis,
    - Dynamic and up-to-date hardware infrastructure,
  - More Economical
    - Organizations no longer need to expend capital upfront for hardware and software purchases
    - Services are provided on a pay-per-use basis,



TPC-H
DECISION-SUPPORT SYSTEM BENCHMARK

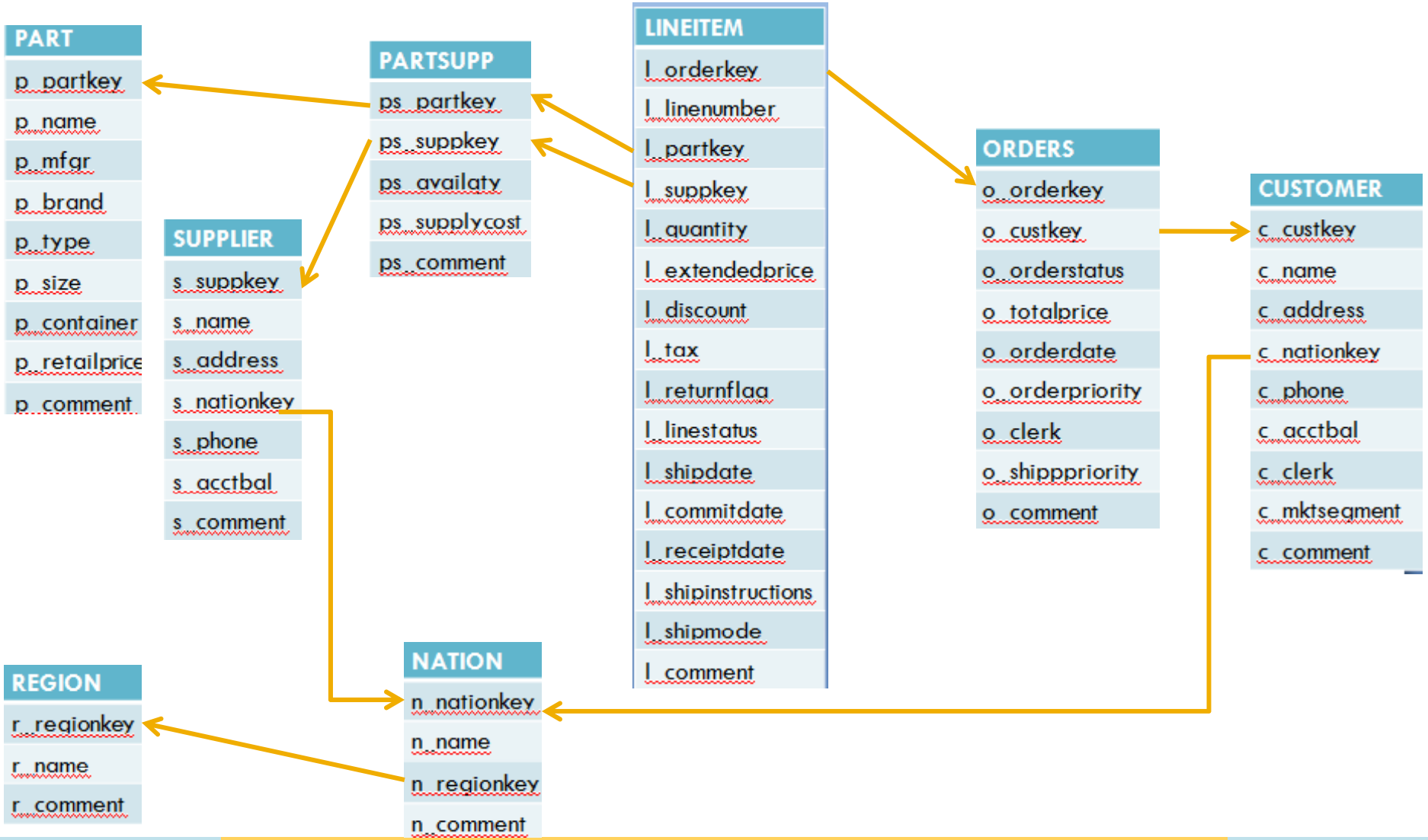
DATA

- Complex DB schema
- Scale factor 1, 10, …, 100,000 correspond respectively to 1GB, 10 GB, …, 100 TB
- 8 data files {lineitem, customer…, region}.tbl
- broad industry-wide relevance

WORKLOAD

- 22 real world business questions
- High degree of complexity
  - Star queries (complex joins)
  - Grouping
  - Nested queries

TPC-H BENCHMARK



HADOOP / PIG LATIN

APACHE HADOOP

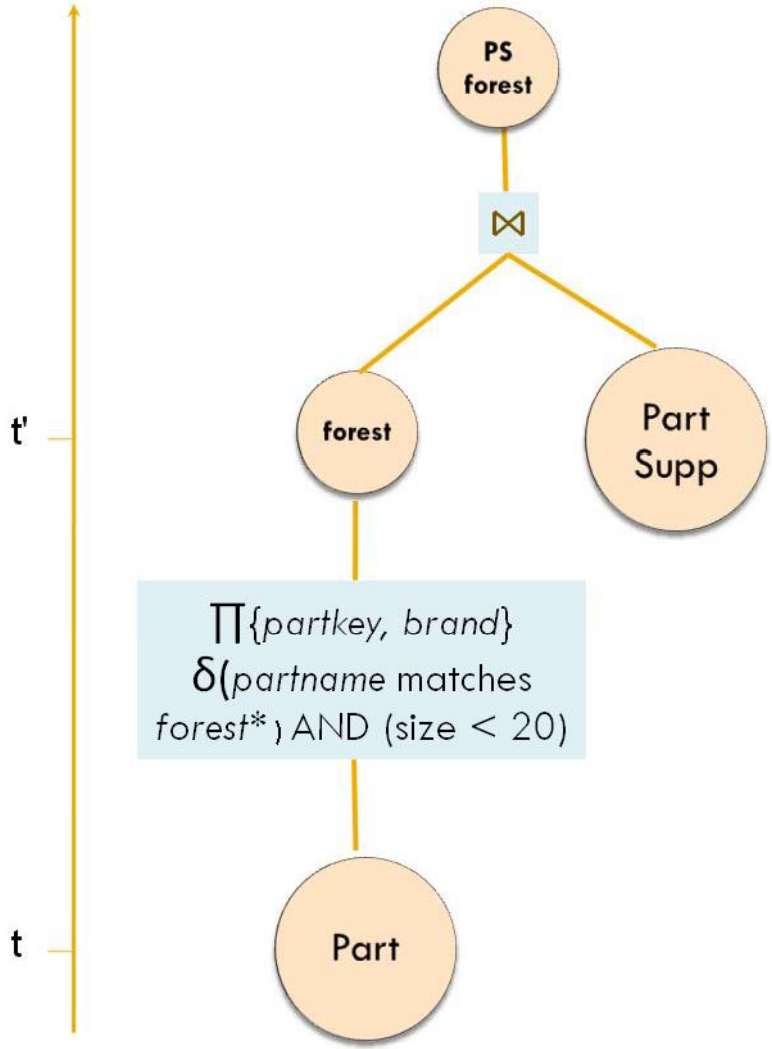
- Framework for running applications on large clusters of commodity hardware.
- Implements computational framework MapReduce
- HDFS: (hadoop distributed file system) stores data on the compute nodes
- Replication & job resoumissions for failures' handling

APACHE PIG LATIN

- high-level language for expressing data analysis programs (filter, projection, join, group, sort, union, …)

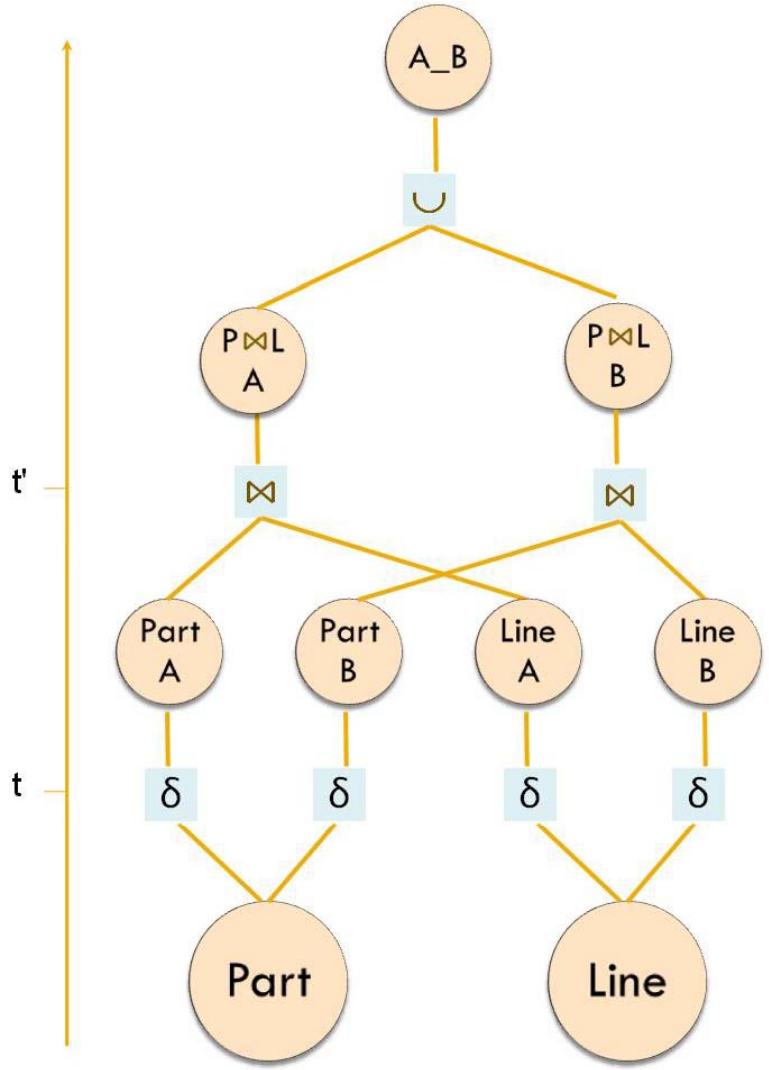
PIG LATIN BENCHMARK
5 TRANSLATION HINTS

1. Load Data for Immediate Processing
   - Better memory management
2. Minimum Relation Scan
   - Conjunction/disjunction of predicates applied once
3. Unary operations prior to binary operations
   - Unary operations (projection, restriction,) reduce data volume

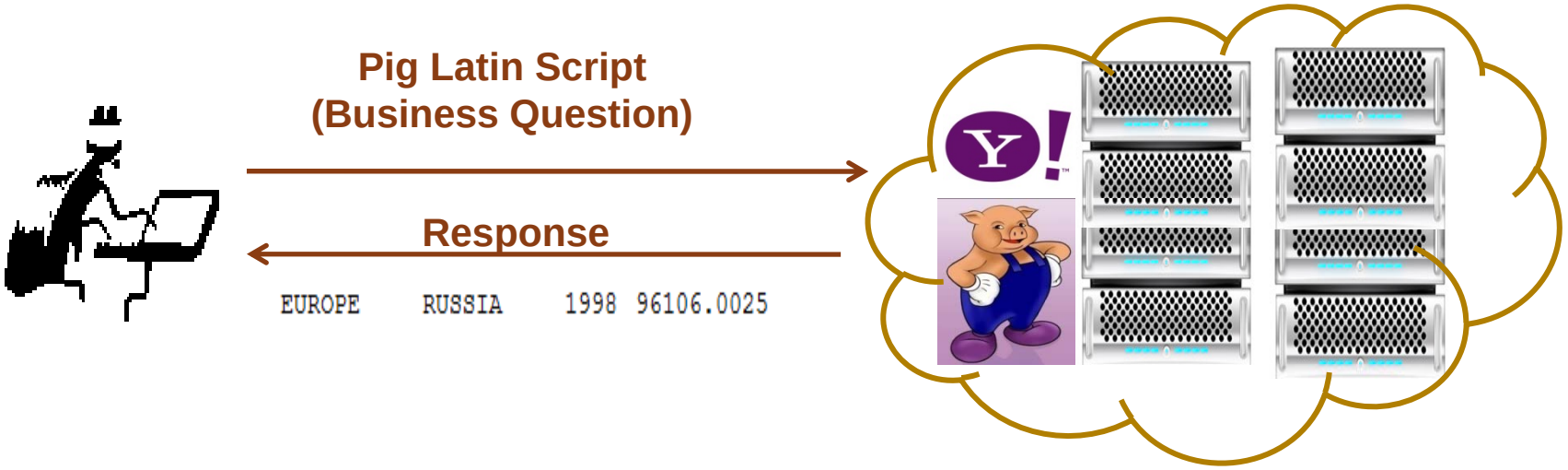


PIG LATIN BENCHMARK
5 TRANSLATION HINTS -CTND

4. Intra-operation parallelism
   - partitioned join
5. Join Algorithm
   - Algorithm
     - hash join,
     - merge join,
   - Star-queries: joins ordering



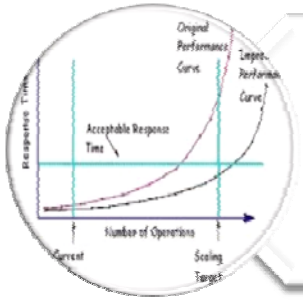
NOMINAL ANALYTICAL SCENARIO



NOMINAL ANALYTICAL SCENARIO

High Cost

- Measured Service, pay as you consume cloud ressources (bandwitdh, CPU, RAM)



Performance Issues

- The same query (with same or different parameters) is executed several times with no optimization

Discontinuity of Service

- Network failure/congestion

COMPLEX!!

How to reduce service cost?

How to improve performances?

How to prevent discontinuity of service?

- Materialized views?
  - Aggregated data replication
- OLAP or not?

Workload Study for Tuning

Cost Management

- Exploit organization hardware resources ?
- Cloud Right size?

TPC-H Workload Numerical Study type A

TPC-H Cube4 *Order Priority Checking*

	Order Priority					
Order Date	-All Order Prioritys	1-URGENT	2-HIGH	3-MEDIUM	4-NOT SPECIFIED	5-LOW
-1993	2 125	389	417	440	436	443
1	500	110	92	101	94	103
2	527	84	117	112	107	107
3	535	93	103	109	102	128
4	563	102	105	118	133	105
+1994	2 126	438	439	420	412	417
+1995	2 022	409	449	396	387	381
+1996	2 103	430	428	412	395	438
+1997	2 090	442	387	395	467	399

$$|order\ date\ \dim| \times |order\ \ priority\ \dim| \times |count\ orders\ measure|$$

$$always\ 135$$

OLAP!
+export to olap server
+MV

TPC-H WORKLOAD NUMERICAL STUDY TYPE B

TPC-H Cube 18 `*Large Volume Orders*`

			Mesures	
Customer Orders	Order Total Price	Order Date	Sum Line QTY	Fact Count
6882	422359.65	1997-04-09	303	7
29158	439687.23	1995-10-21	305	7

$$|order\ \dim| \times |sum\ line\ qty\ measure|$$

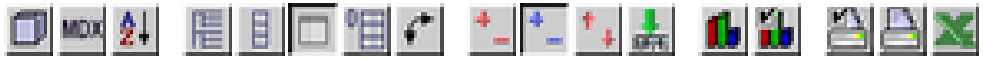
$$SF \times 1{,}500{,}000$$

$$almost\ 3.8\ ppm\ of\ \ orders\ have \sum line\ qty > 300,\ for\ SF = 1$$

Not OLAP!
+MV

TPC-H WORKLOAD NUMERICAL STUDY TYPE C

TPC-H Cube 2 `*Minimum Cost Supplier*`



Supplier	Supp Acct Bal	Supp Phone	Supp Comment	Supp Address
+AFRICA				
+AMERICA				
+ASIA				
+EUROPE				
-MIDDLE EAST				
+EGYPT				
+IRAN				
-IRAQ				
Supplier#000000005	-283.84	21-151-690-3663	. slyly regular pinto bea	Gcdm2rJRzl5qlTVzc

Part	Part MFGR	Part Size	Part Type	Mesures: Min Supply Cost
All Parts				16,82
1	Manufacturer#1	7	PROMO BURNISHED COPPER	
2	Manufacturer#1	1	LARGE BRUSHED BRASS	
3	Manufacturer#4	21	STANDARD POLISHED BRASS	
4	Manufacturer#3	14	SMALL PLATED BRASS	113,97

$$|\text{supplier dim}| \times |\text{part dim}| \times |\text{min supply cost } \textit{measure}|$$

OLAP!
MV storage cost

$$SF^2 \times 2{,}000{,}000{,}000$$

best supplier in each region for each part!

TPC-H Workload Numerical Study

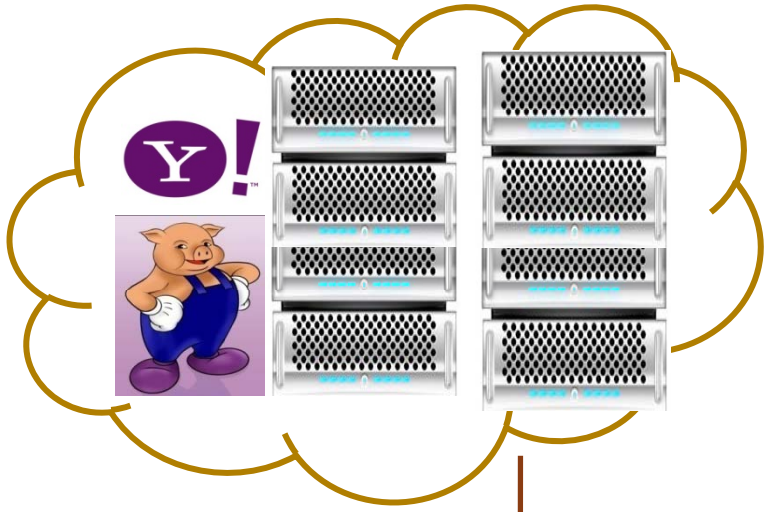
Type	Features	TPC-H Business Questions (OLAP Cube)
A	• Medium dimensionality • Result is TPC-H Scale Factor independent	Q1, Q3, Q4, Q5, Q6, Q7, Q8, Q12, Q13, Q14, Q16, Q19, Q22 13 business questions
B	• High dimensionality • few results, lots of empty cells	Q15, Q18 2 business questions
C	• High dimensionality • Result % of Scale Factor	Q2, Q9, Q10, Q11, Q17, Q20, Q21 7 business questions

CLOUD COST MANAGEMENT

- Measured Service
  - pay as you go
  - CPU + Memory + Bandwidth
- "*When users understand the relationship between cost and consumption, everybody wins*" –Ron Miller
- **Emerging need to understand, manage and proactively control costs across the cloud**
  - Resource Utilization Monitoring
  - Right size w.r.t. both performances & cost (client and provider)
    - Green cloud through energy saving

BETTER SCENARIO

Pig Latin Script (Generalized Business Question)



Interaction

Pre-aggregated Data

```
MIDDLE EAST      IRAN  1993  255735.04959999997
MIDDLE EAST      IRAN  1992  203817.70789999998
MIDDLE EAST      EGYPT 1998  387012.1518
MIDDLE EAST      EGYPT 1997  779865.8600999999
MIDDLE EAST      EGYPT 1996  643782.6547000001
MIDDLE EAST      EGYPT 1995  881419.7011
MIDDLE EAST      EGYPT 1994  1339454.0117
MIDDLE EAST      EGYPT 1993  803573.9572000001
MIDDLE EAST      EGYPT 1992  705541.2589
EUROPE      UNITED KINGDOM   1998  205051.5488
EUROPE      UNITED KINGDOM   1997  413469.2654000
EUROPE      UNITED KINGDOM   1996  291845.8779
EUROPE      UNITED KINGDOM   1995  647750.816
EUROPE      UNITED KINGDOM   1994  512853.7625
EUROPE      UNITED KINGDOM   1993  146581.1697999
EUROPE      UNITED KINGDOM   1992  305019.6303
EUROPE      RUSSIA      1998  96106.0025
EUROPE      RUSSIA      1997  370882.6169
```

OLAP Client

						Measures		
Region	My Image	P	P0	P1	P2	Measures[0]	Measures[1]	Measures[2]
-All Region[0]		0	V00			856.44	820.95	983.55
+Region[0]		0		V10		1,095.05	1,087.79	958.23
-Region[1]		1		V11		1,052.05	1,107.59	980.32
-City[0]	✓	0			V20	1,088.31	884.90	1,085.73
+City[1]	✓	1			V21	1,336.43	943.27	1,117.62
+City[2]	✓	2			V22	969.34	1,086.70	1,094.71
-City[3]	✓	3			V23	964.30	900.58	953.65
+City[4]	✓	4			V24	915.59	1,032.98	876.19
+City[5]	✓	5			V25	1,023.12	1,242.39	988.85
-City[6]	✓	6			V26	949.16	941.62	1,077.48
+City[7]	✓	7			V27	1,055.92	1,095.56	1,078.26
+Region[2]		2		V12		1,027.36	1,084.88	1,009.71
+Region[3]		3		V13		972.10	932.27	836.34
+Region[4]		4		V14		919.44	926.47	873.20

Import Data into an on-site OLAP server

PERFORMANCE MEASUREMENTS

G5K

French GRID platform: a **large scale nation wide infrastructure for Grid research.**

- Bordeaux Site
  - Borderel: 24GB RAM, 4 AMD CPUs, 2.27 GHz, and 4cores/CPU.
  - Borderline: : 32GB RAM, 4 Intel Xeon CPUs, 2.6 GHz, and 2 cores/CPU.
  - Ethernet10Gbps

TPC-H

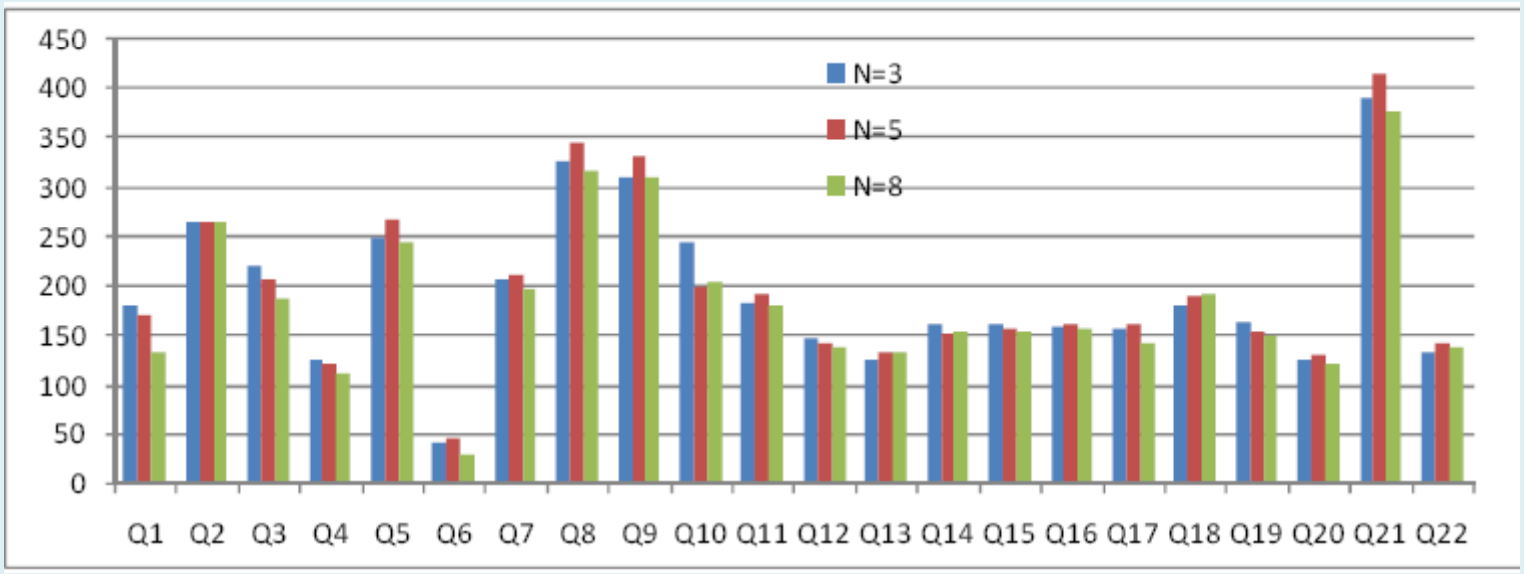
- TPC-H Benchmark
- **SF=1**
  - 1.1GB source files
  - 4.5GB single big file
- **SF=10**
  - 11GB source files
  - 45GB single big file

Pig/HDFS

- **Apache Hadoop 0.20**
  - N=3, 5 or 8
  - one Hadoop Master
  - (2, 4 or 7) Workers
- **Apache Pig 0.8.1**

PERFORMANCE MEASUREMENTS

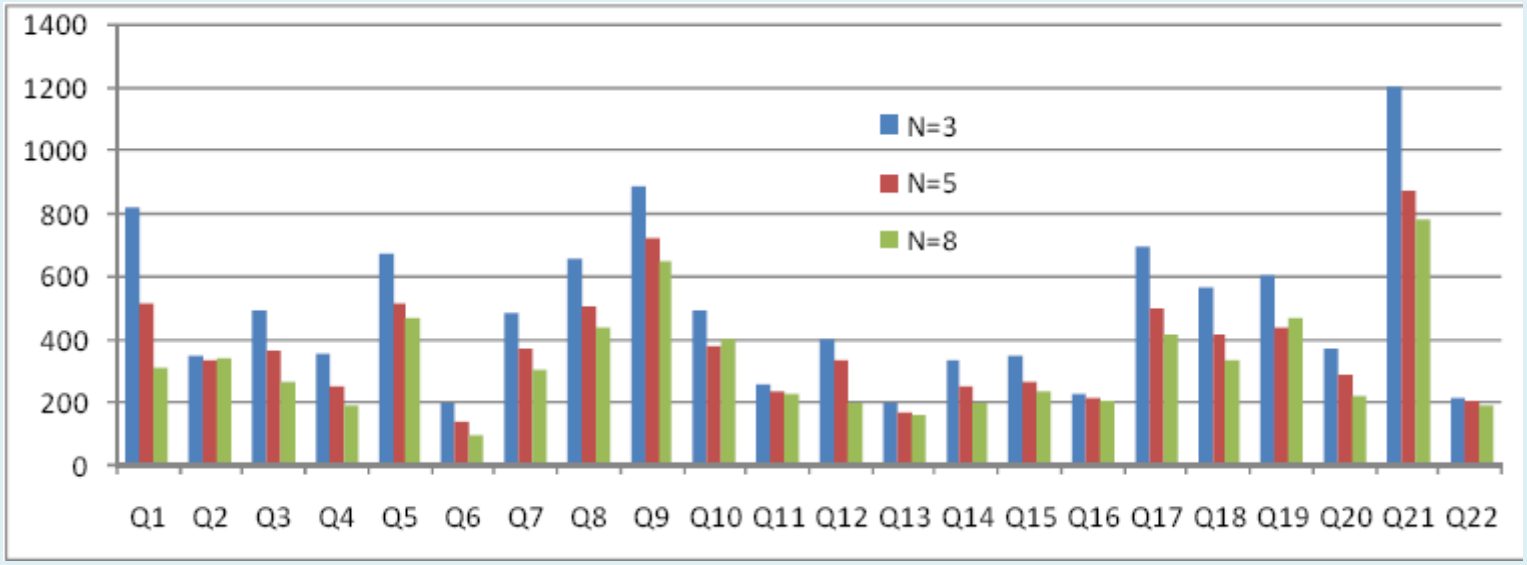
Original TPC-H 1GB	Original TPC-H 11GB	Original 1.1 vs 11GB	Big File	Big File 4.5GB	Big File 45GB



- Except business questions which do not perform join operations: No improvement when cluster size doubles

PERFORMANCE MEASUREMENTS

Original TPC-H 1GB	**Original TPC-H 11GB**	Original 1.1 vs 11GB	Big File	Big File 4.5GB	Big File 45GB



- Improvement when cluster size doubles
- More data so it's nice to have more storage & computing nodes

PERFORMANCE MEASUREMENTS

Elapsed times for SF=10 (11GB) are

- At maximum 5 times elapsed times obtained for SF=1 (1.1GB)
- In average twice elapsed times obtained for SF=1 (1.1GB)

PERFORMANCE MEASUREMENTS

Joining partitionned files is complex!

Combine all files into one file

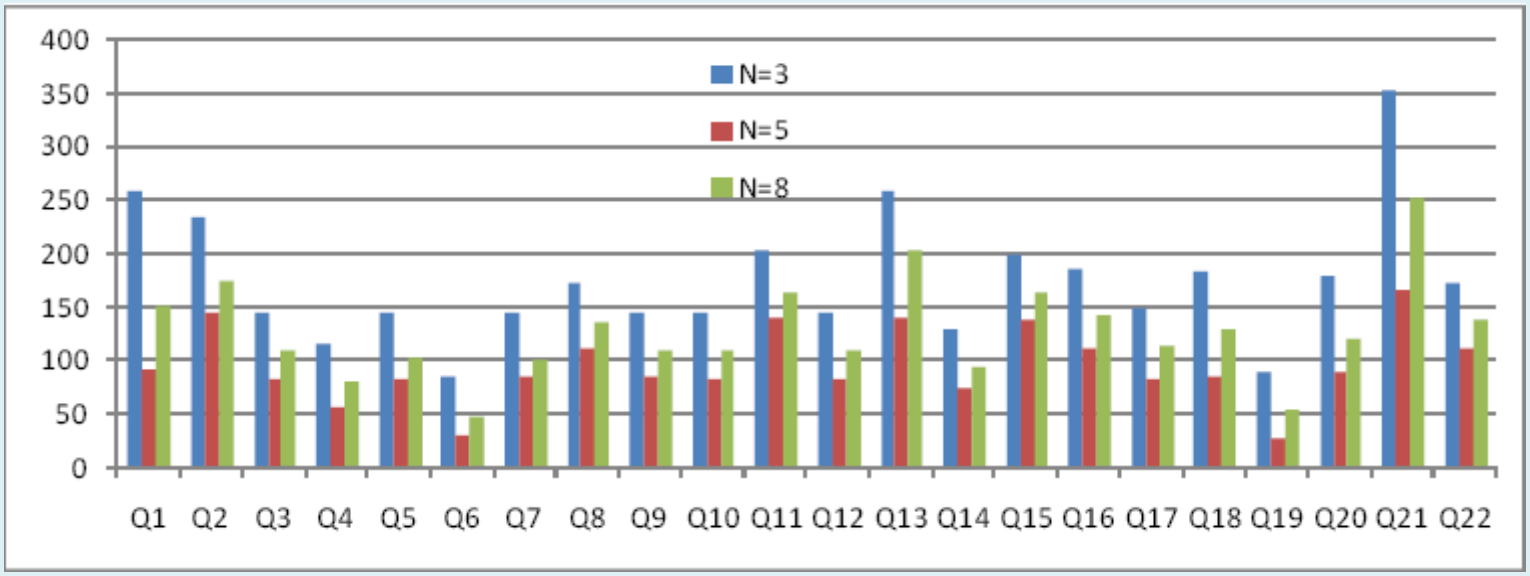
SF=1 → 4.5GB

SF=10 → 45GB

- Evaluation of Pig/MR without joins
- Denormalization
  - saves join cost
  - increases required storage space (≈ ×4 for TPC-H)

PERFORMANCE MEASUREMENTS

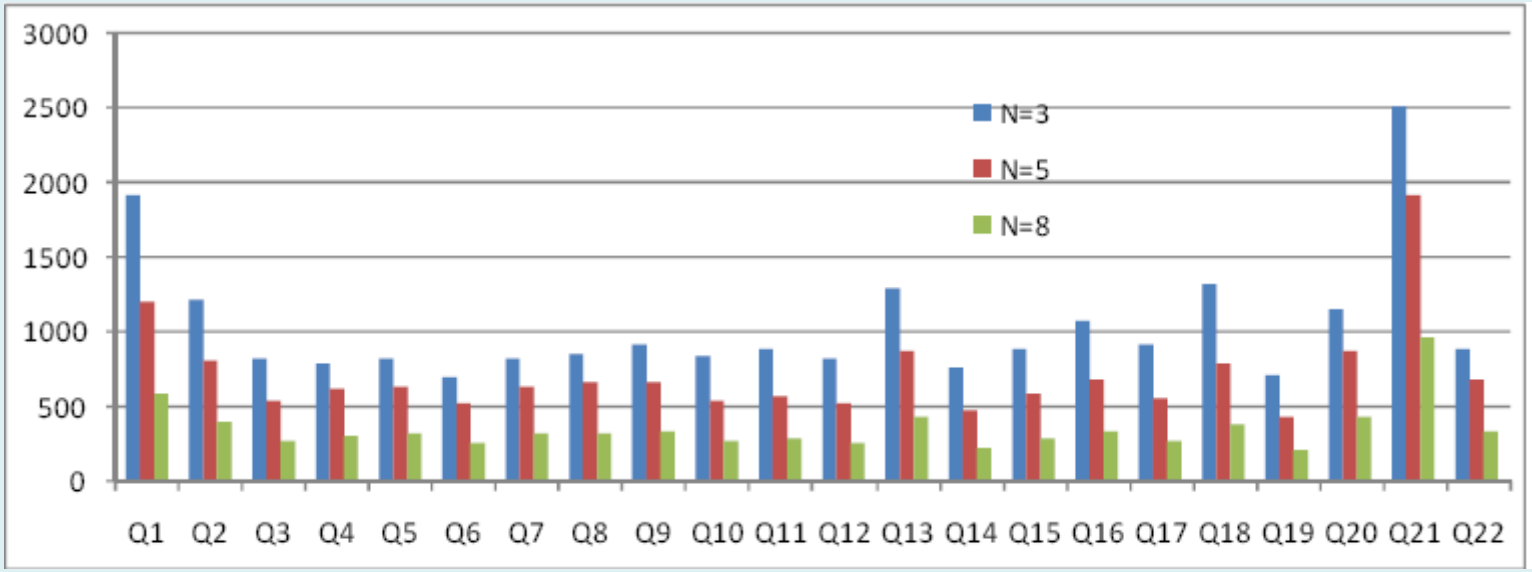
Original TPC-H 1GB	Original TPC-H 11GB	Original 1.1 vs 11GB	Big File	**Big File 4.5GB**	Big File 45GB



- Compared to (SF=1, 1.1GB), improvements range from 10% to 80%,
- performance degradation with more than 4 workers (N=5): this is due to MR framework (before reduce phase, data is grouped and sorted which has a cost when involving more storage and computing nodes)

PERFORMANCE MEASUREMENTS

Original TPC-H 1GB	Original TPC-H 11GB	Original 1.1 vs 11GB	Big File	Big File 4.5GB	**Big File 45GB**



- Compared to (SF=10, 11GB), after affording more nodes we obtain similar results
- Compared to (SF=1, 4.5GB), elapsed times are less than 10× for same cluster size
- Performance degradation for queries which do not perform joins

  Q1 executes over 7GB lineitem file (SF=10,11GB), now it executes over 45GB file

PERFORMANCE MEASUREMENTS

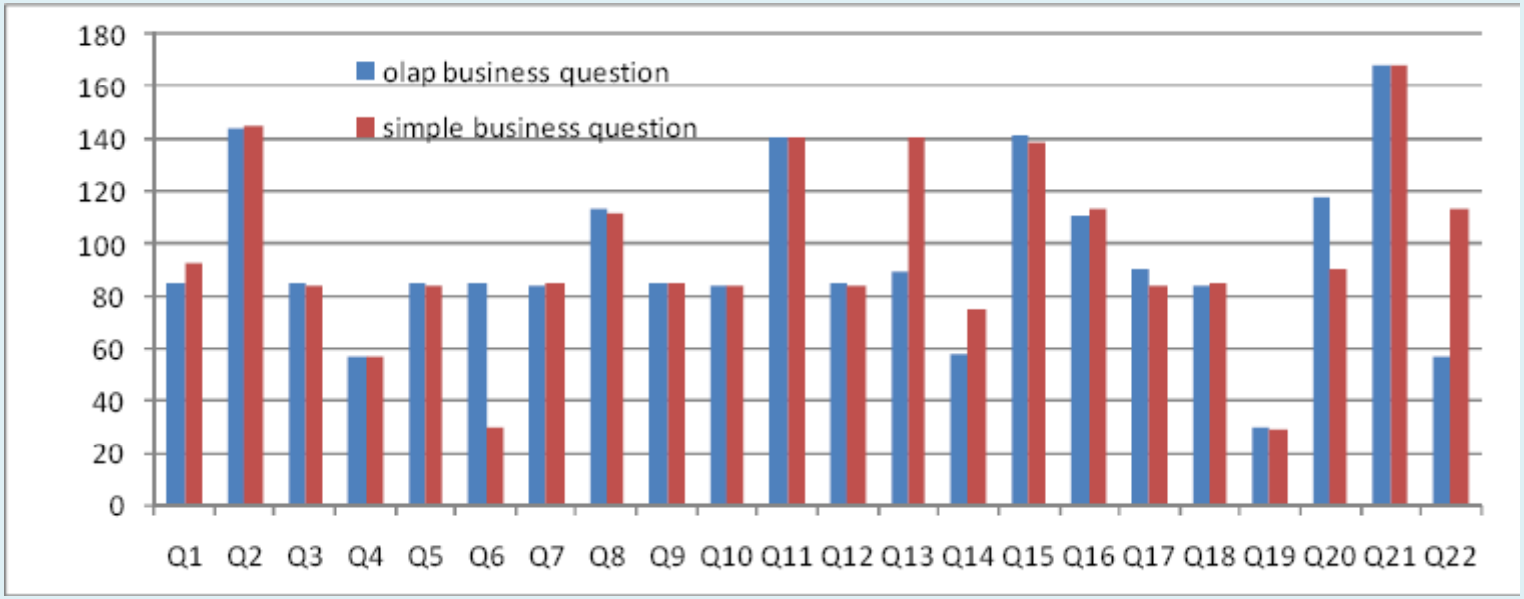
Aggregated data

- TPC-H business questions type A (SF independent & small resultset)
- TPC-H business questions type B (very very small resultset)
- 15 business questions from 22

Tradeoff between space & computation

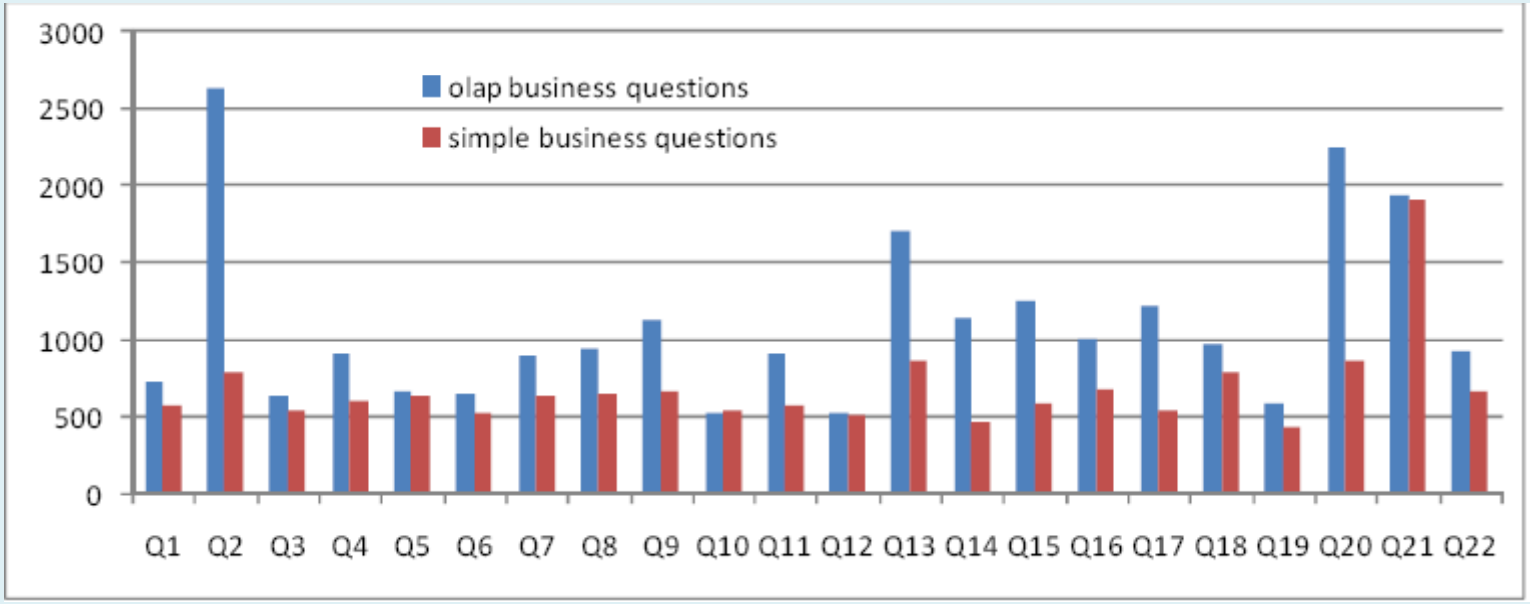
- TPC-H business questions type C
- Add *derived fields*
  - Q2: check (true) minimum supplycost by supplier for a part in PartSupp
  - Q17: average_line_quantity field for each part
  - Q20: sum_lines_quantities_per_year for each supplier
  - Q21: number of waiting orders for each supplier
- 7 business questions from 22

PERFORMANCE MEASUREMENTS



- Average degradation is 5%
- done once + time to retreive data

PERFORMANCE MEASUREMENTS



- Average degradation is 60%
- done once + time to retreive data

RELATED WORK

- Implementation of relational operations using MR framework
  - Kim et al. –*MRBench, 2008*
  - Nominal analytics scenario
  - TPC-H benchmarking for SF=1,3
- Translation from SQL to Pig Latin
  - Iu et al. –*Hadoop to SQL, 2010*
  - Lee et al. –*Ysmart, 2011*
- Other pig latin use cases
  - Shatzle et al, RDF data, 2011
  - Loebman etal. Astrophysical data, 2009

CONCLUSION

- TPC-H in-depth numerical study
- OLAP in the cloud
  - Scenarios
  - Implementation
    - Pig / Hadoop Distributed File System
  - Performances
    - Various cluster sizes
    - Various data volumes
    - Various schemas (with and without joins)

FUTURE WORK
AQP

- Approximate Query Processing in clouds
  - Most Distributed File Systems implement replication for high availability
  - MDS erasure codes outperform replication from two perspectives (i) storage cost and (ii) minimal operation cost of redundant data management
    - New Hadoop release
    - Facebook
  - → Generalized framework for approximate data analytics in the cloud coping with nodes' failure

FUTURE WORK
PIG LATIN++

- Most of TPC-H business questions scripts are composed of jobs, which execute sequentially,

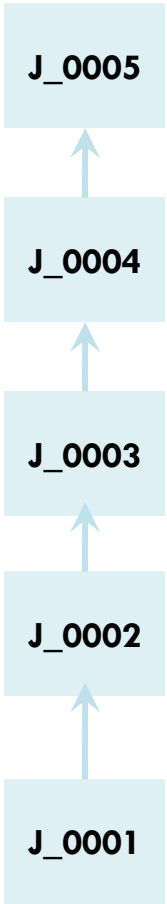
  → some branches of the DAG are unnecessarily blocked!

  - namely Q1, Q3, Q4, Q9, Q10, Q11, Q12, Q13, Q14, Q16, Q17 and Q18

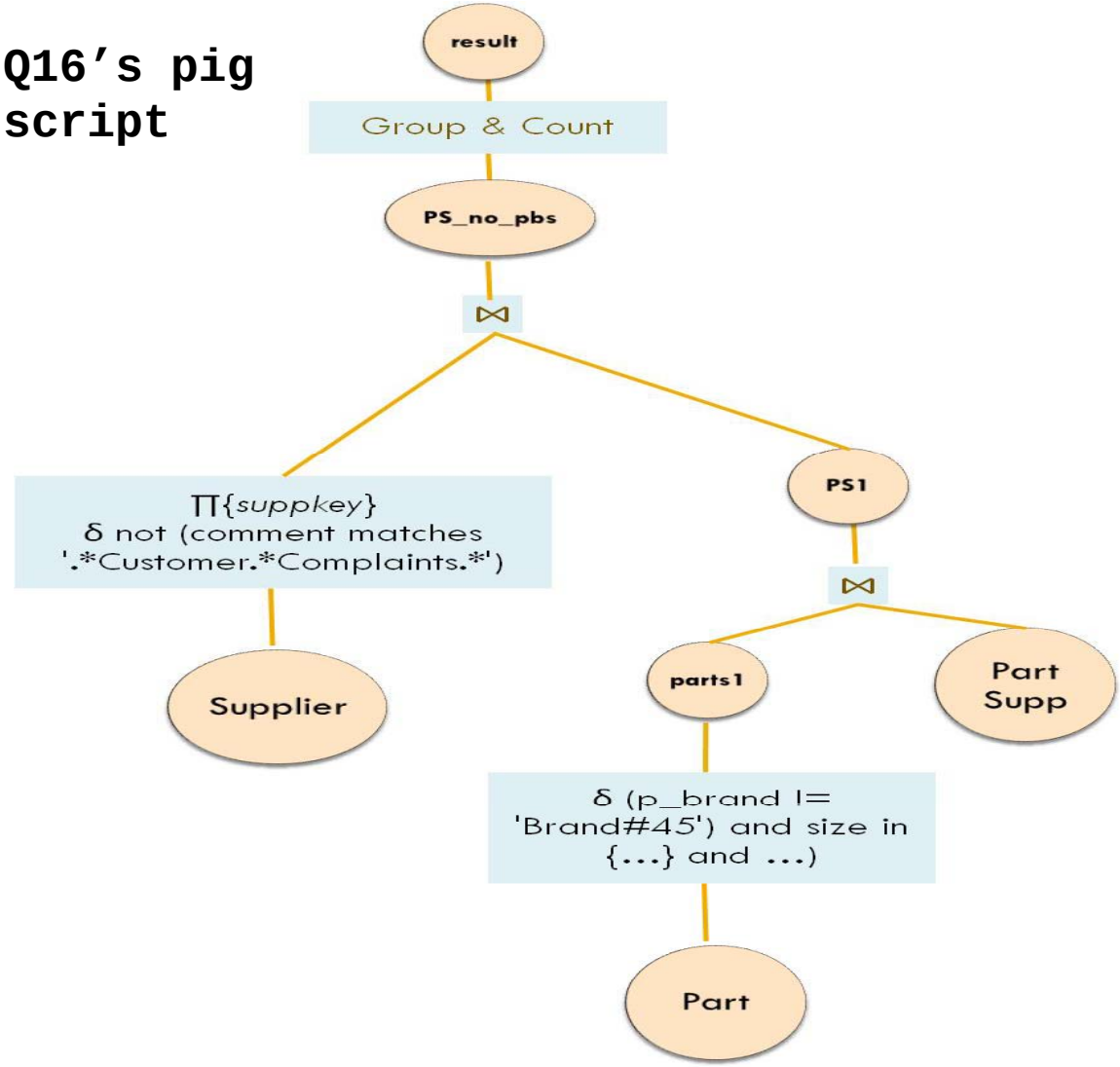
- Pig Latin Enhancements
  - Investigate intra-operation Parallelism for better performances of Pig Scripts
  - Investigate better job definitions strategies, in order to increase inter-job parallelism

FUTURE WORK
PIG LATIN++⟩⟩ Q16 EXPLE

Q16's DAG



Q16's pig script



TPC-H Analtycs Scenarios and Performances on Hadoop Data Clouds

Thank you for your Attention

Q & A

?

24th, Apr. 2012

4th International. Conference on Networked Digital Technologies

NDT'12.Dubai. UAE