An Investigation into the Efficiency of the Star Schema

Bill Qualls

DePaul University

IS549 - Data Warehousing and Data Mining

Instructor: Dr. Lu Zhang

November 9, 2010

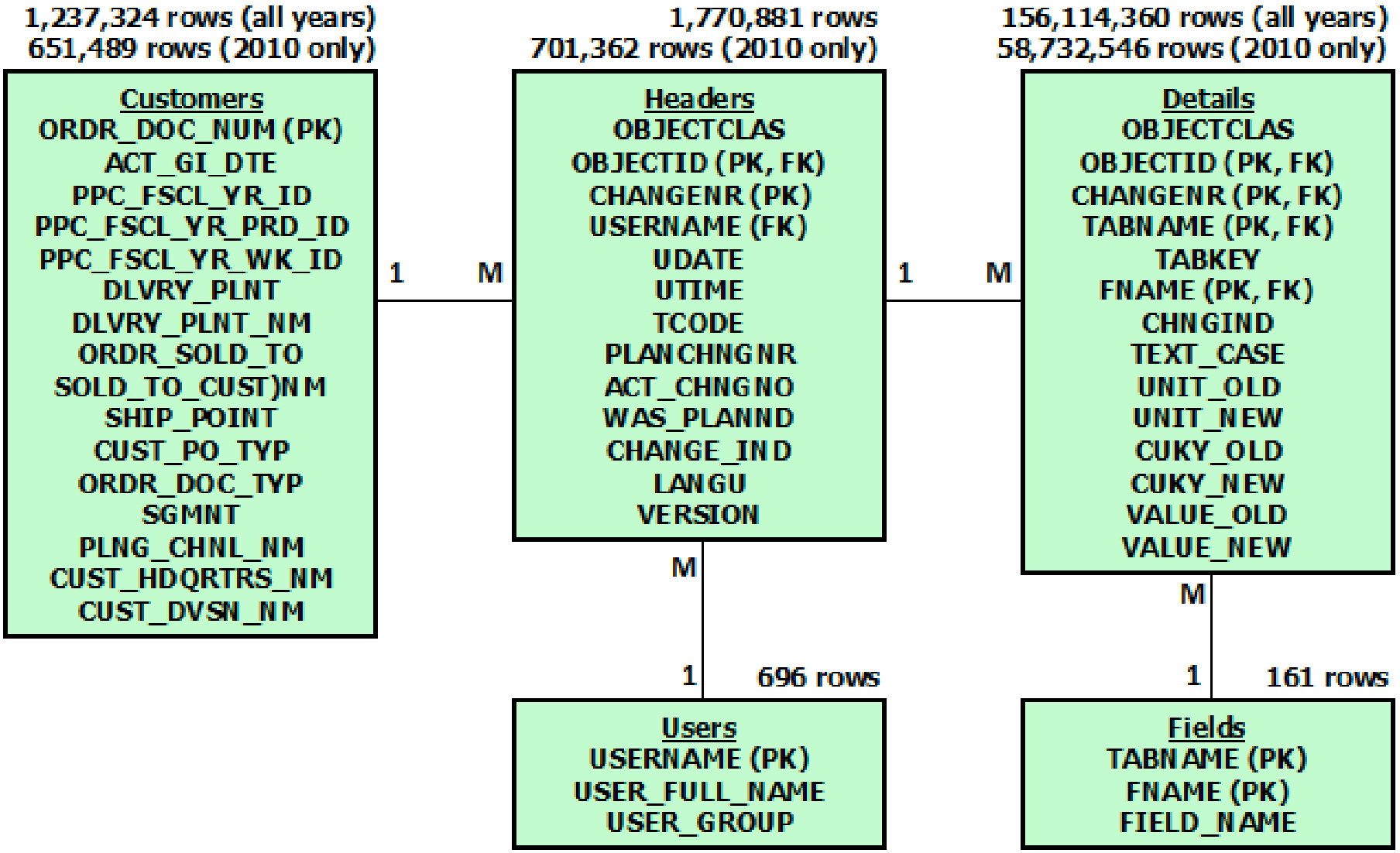
Introduction

- 3NF supports OLTP
- Star schema supports OLAP.
- Oh really?
- I had to see for myself....

Coincidentally...

- I received a work assignment requiring queries on tables with tens of millions of rows.
- I hoped to find -- and present -- empirical rather than anecdotal evidence that query times are, indeed, reduced with the star schema.
- Because I worked for an analytics company which is partnered with SAS, Inc., all coding was done with SAS®.

As Received...Normalized (almost)

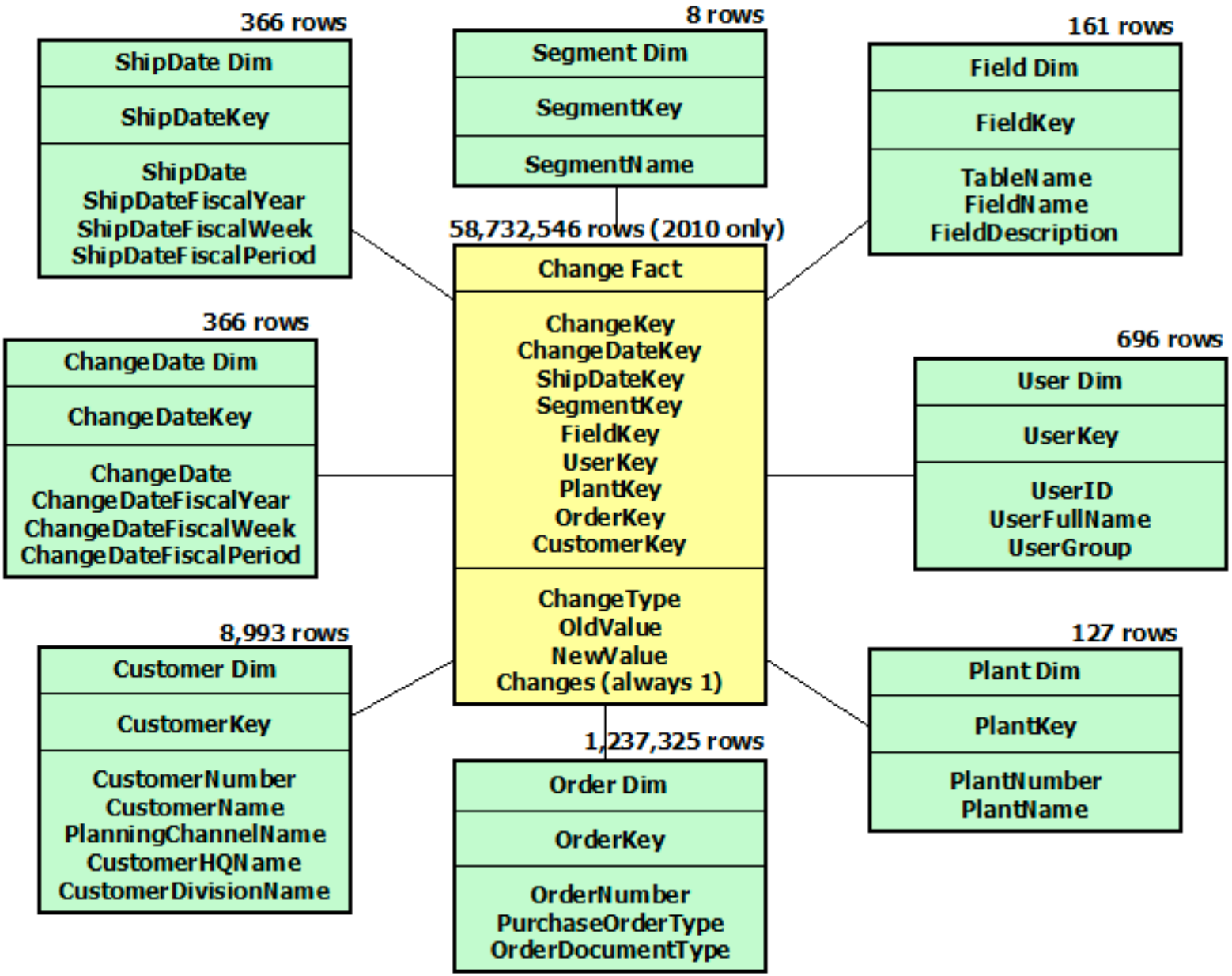


I also created one big table: "Joined"

58,732,546 rows (2010 only)

Changes
OBJECTID
CHANGENR
TABNAME
TABKEY
FNAME
CHNGIND
VALUE_OLD
VALUE_NEW
USERNAME
UDATE
PLANCHNGNR
ACT_GI_DTE
DLVRY_PLNT
DLVRY_PLNT_NM
SOLD_TO_CUST_NM
CUST_PO_TYP
ORDR_DOC_TYP
SGMNT
PLNG_CHNL_NM
CUST_HDQRTRS_NM
CUST_DVSN_NM
USER_FULL_NAME
USER_GROUP

Star Schema


366 rows
ShipDate Dim
ShipDateKey
ShipDate
ShipDateFiscalYear
ShipDateFiscalWeek
ShipDateFiscalPeriod
8 rows
Segment Dim
SegmentKey
SegmentName
161 rows
Field Dim
FieldKey
TableName
FieldName
FieldDescription
58,732,546 rows (2010 only)
Change Fact
ChangeKey
ChangeDateKey
ShipDateKey
SegmentKey
FieldKey
UserKey
PlantKey
OrderKey
CustomerKey
ChangeType
OldValue
NewValue
Changes (always 1)
366 rows
ChangeDate Dim
ChangeDateKey
ChangeDate
ChangeDateFiscalYear
ChangeDateFiscalWeek
ChangeDateFiscalPeriod
696 rows
User Dim
UserKey
UserID
UserFullName
UserGroup
8,993 rows
Customer Dim
CustomerKey
CustomerNumber
CustomerName
PlanningChannelName
CustomerHQName
CustomerDivisionName
127 rows
Plant Dim
PlantKey
PlantNumber
PlantName
1,237,325 rows
Order Dim
OrderKey
OrderNumber
PurchaseOrderType
OrderDocumentType

Test Queries

#1 – Show the total number of changes by Type.

#2 – Show the total number of changes by Customer.

#3 – Show the total number of changes to the Order Quantity field by Customer by days out from the ship date.

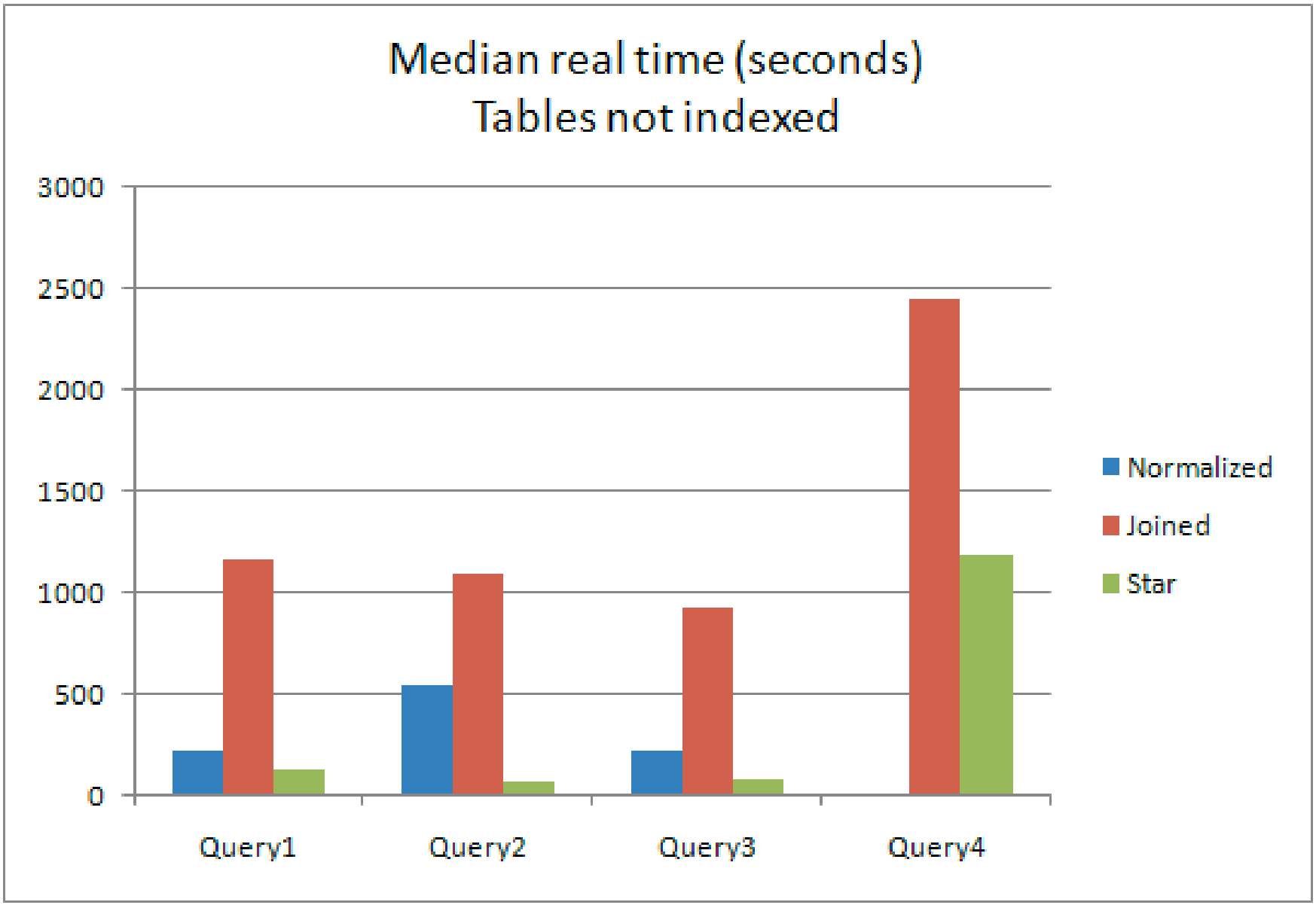
#4 – Show the total number of changes by Customer / Type / User / Field.

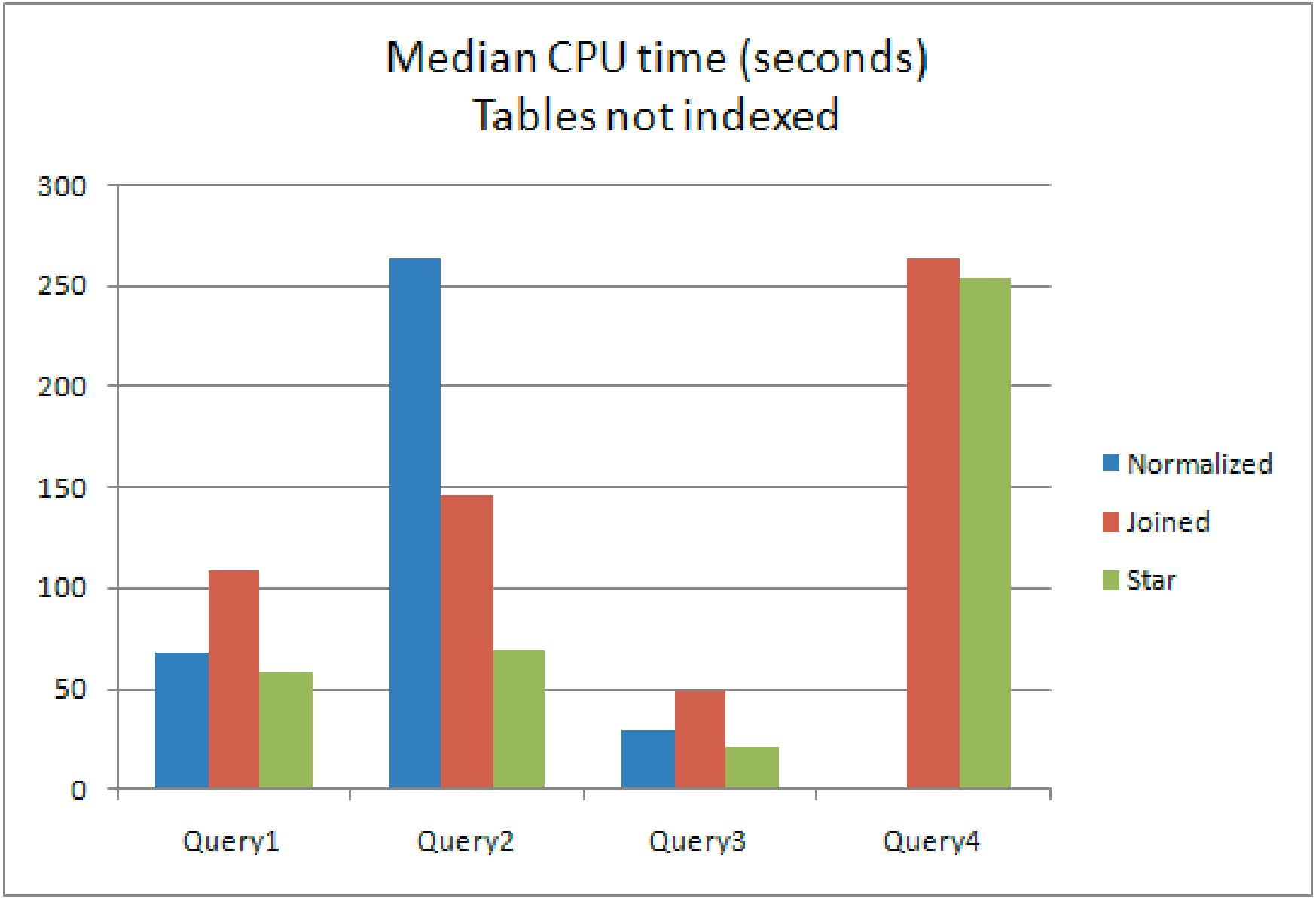
Query Results - Tables NOT Indexed

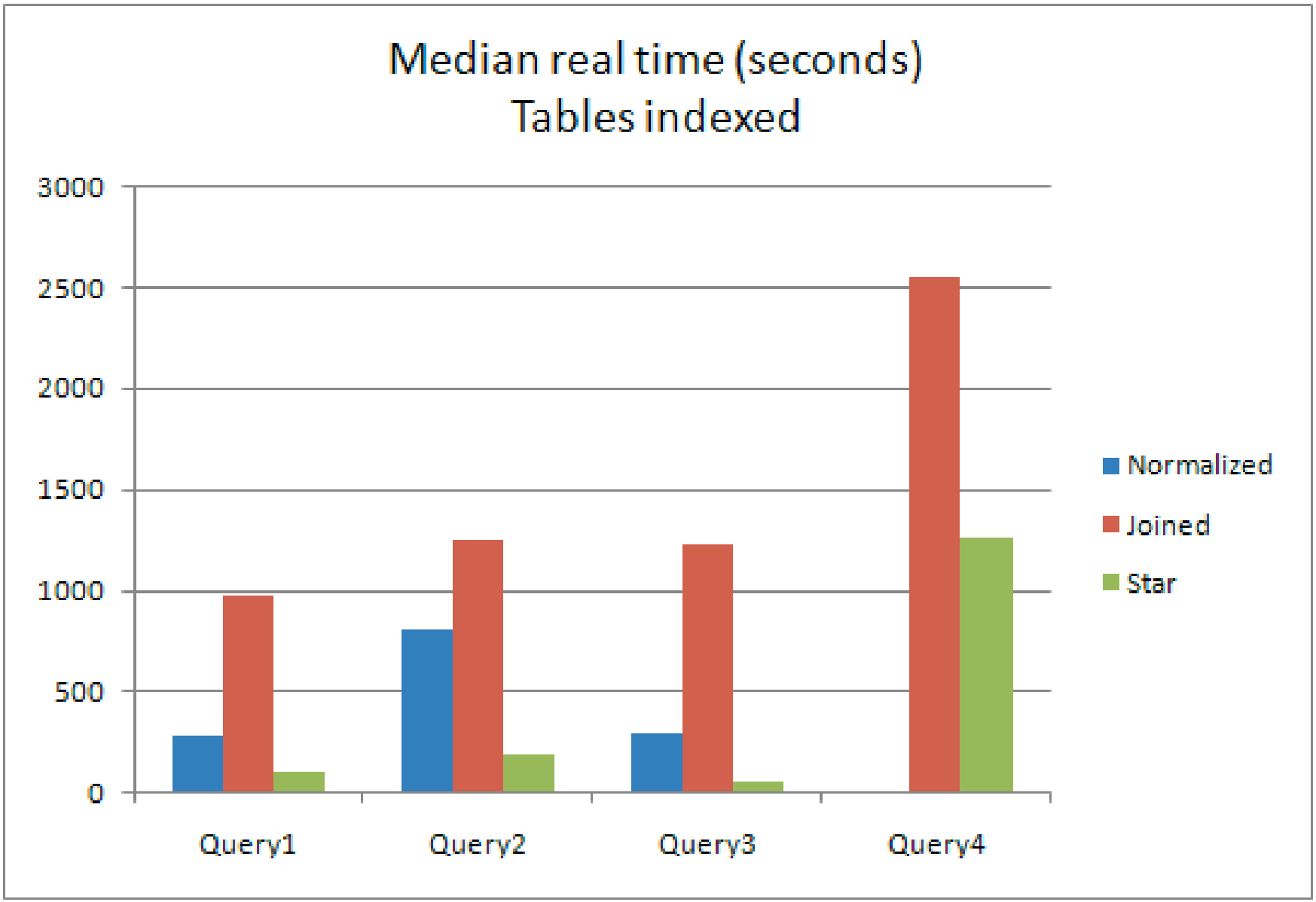
Query	1	2	3	4
Not Indexed				
Normalized	4:48, 1:12	11:50, 5:09	3:39, 0:30	failed
	3:41, 1:08	8:40, 4:22	3:50, 0:29	failed
	3:35, 1:08	8:59, 4:24	3:27, 0:30	failed
Joined	20:26, 1:49	18:15, 2:26	15:22, 0:49	40:51, 4:24
	19:27, 2:45	18:40, 2:26	15:03, 0:55	40:19, 4:21
	16:24, 1:46	17:52, 2:16	20:12, 0:46	45:26, 4:31
Star	2:03, 0:58	2:59, 1:48	1:37, 0:35	26:20, 4:14
	2:06, 0:57	1:04, 1:09	1:08, 0:21	19:49, 4:51
	2:10, 0:59	1:06, 1:09	1:06, 0:20	17:50, 4:11

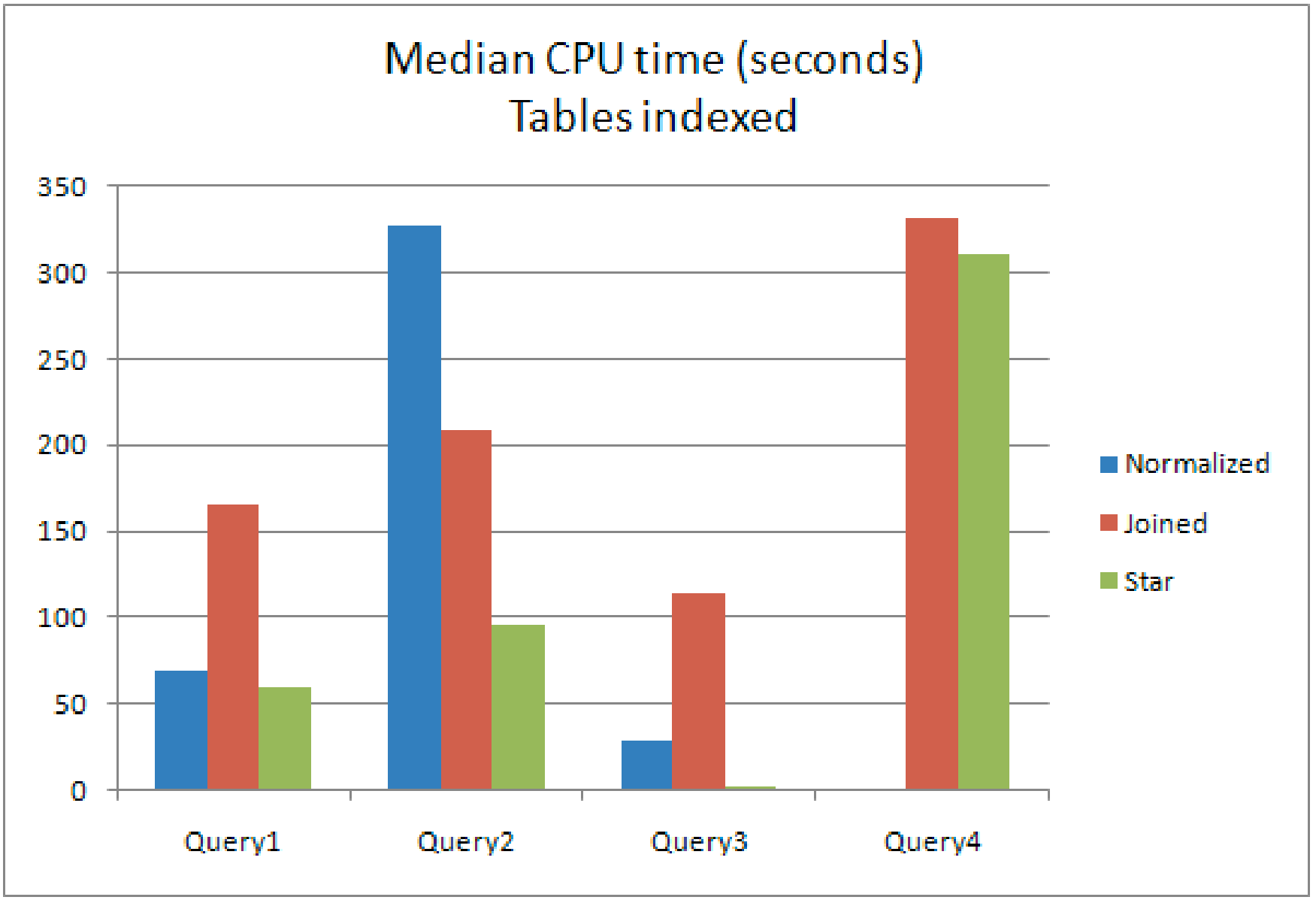
Query Results - Tables Indexed

Query	1	2	3	4
Indexed				
Normalized	6:28, 1:33	13:28, 5:27	4:43, 0:32	failed
	4:47, 1:07	13:03, 5:27	4:55, 0:28	failed
	4:48, 1:09	15:18, 5:34	5:04, 0:29	failed
Joined	15:23, 2:43	20:47, 3:28	21:41, 1:54	43:57, 5:39
	16:40, 2:50	22:05, 3:32	20:00, 1:51	42:28, 5:32
	16:13, 2:46	18:15, 3:22	20:29, 1:55	42:39, 5:29
Star	3:08, 1:13	2:09, 1:25	0:28, 0:02	26:07, 5:10
	1:38, 0:58	3:35, 1:36	0:54, 0:03	19:34, 4:54
	1:43, 1:00	3:11, 1:44	0:52, 0:03	21:02, 5:11


Median real time (seconds)
Tables not indexed
3000
2500
2000
1500
1000
500
0
Query1
Query2
Query3
Query4
Normalized
Joined
Star


Median CPU time (seconds)
Tables not indexed
300
250
200
150
100
50
0
Query1
Query2
Query3
Query4
Normalized
Joined
Star


Median real time (seconds)
Tables indexed
3000
2500
2000
1500
1000
500
0
Query1
Query2
Query3
Query4
Normalized
Joined
Star


Median CPU time (seconds)
Tables indexed
350
300
250
200
150
100
50
0
Query1
Query2
Query3
Query4
Normalized
Joined
Star

Analysis

- The star schema queries executed faster than their normalized and joined counterparts, in every query, both in terms of real time and CPU time.
- Curious, but not directly related to the topic at hand, is that in most cases, regardless of the schema used, the queries ran faster when the tables were not indexed! Further investigation is warranted.

Conclusion

- The purpose of this paper was to provide empirical data about the relative efficiency of the star schema.
- This paper has demonstrated that the use of the star schema can significantly reduce query response times using SAS®.
- It would be interesting to see the results of the same queries with the same data running under a true DBMS such as Oracle.