



Product Deck

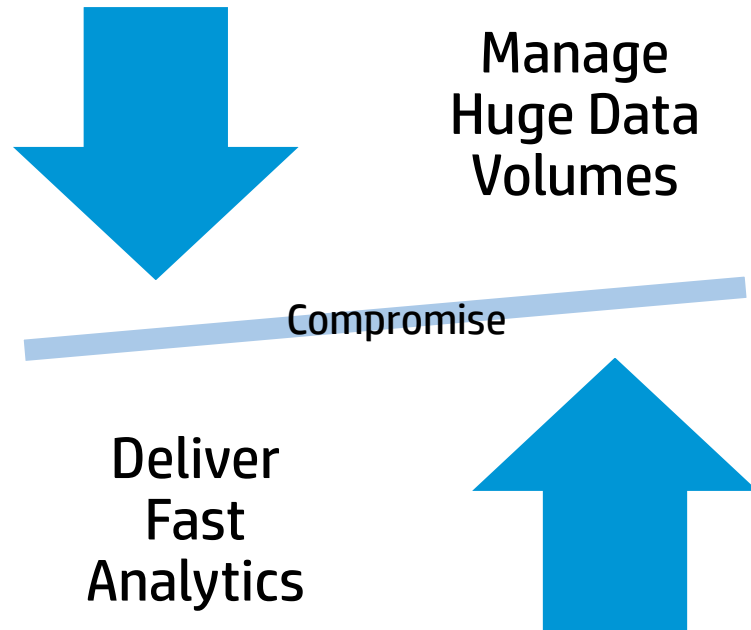
Big Data solutions are full of compromise

Traditional Enterprise Databases
















- The original SQL databases did not envision today's data volumes
- Vendors scrambling to handle bigger data volumes by tacking on Hadoop technologies and retrofitting legacy technologies

Hadoop-Based Solutions

- Major Hadoop vendors strive to meet the standard with SQL on Hadoop
 - NoSQL is incomplete SQL
 - Analytics Performance is Limited
 - Not a substitute for a full implementation of SQL



Big Data Platform Compromises

Key Capability	HP Vertica	Legacy Analytics Platforms	Hadoop Analytics Platforms	Use Case
Ingest Huge Data Volumes				Petabyte of data
Deliver Fast Analytics				What-if, A/B testing
Work with Legacy Tools				SQL-based Visualization ETL
Support Data Scientists				Python and R
Advanced Analytics				Joins, Complex Data Types SQL-based Predictive Analytics

HP Vertica Core Capabilities

Solutions and Applications

Data Sources & Formats

Advanced Analytics

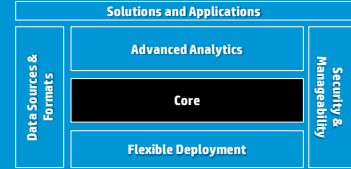
Core

Flexible Deployment

Security & Manageability



Core Capabilities



Core Capabilities – Built for Speed

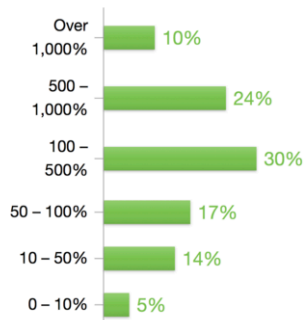
We boost performance




Research by  TechValidate

HP Vertica means insanely high query performance!

By what % have you increased query performance through improvement in response time with HP Vertica? Please provide your best estimate.



Source:  Survey of 212 users of HP Software Vertica

Published: Feb. 24, 2014

TVID: F4B-E32-A1C

What 1000% means:

Use to take	Now takes
1 hour	3.6 Seconds
8 hours (overnight)	Under 30 seconds

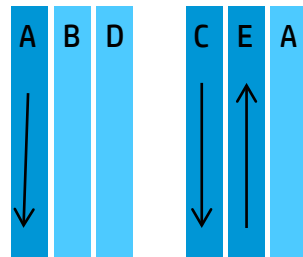
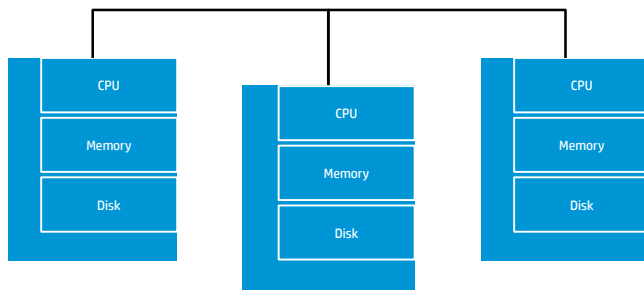
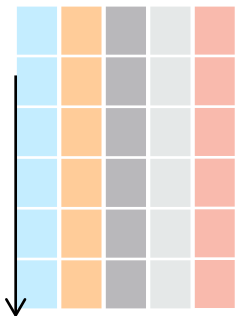
"When we did the first queries, they were done so fast, we thought they were broken."

- Michael Relich, Guess?



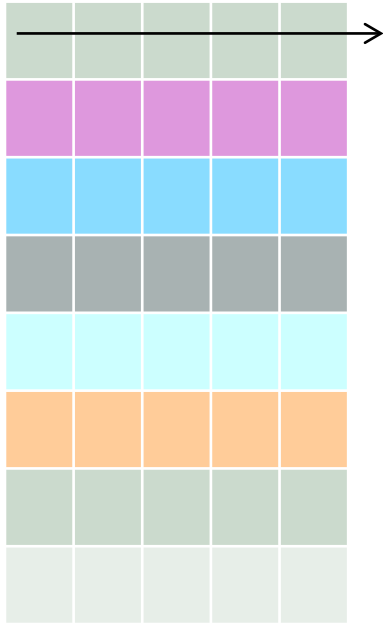
Secrets to Achieving Performance Increases

Columnar Storage	Compression	MPP Scale-Out	Distributed Query	Projections
Speeds Query Time by Reading Only Necessary Data	Lowers costly I/O to boost overall performance	Provides high scalability on clusters with no name node or other single point of failure	Any node can initiate the queries and use other nodes for work. No single point of failure	Combine high availability with special optimizations for query performance



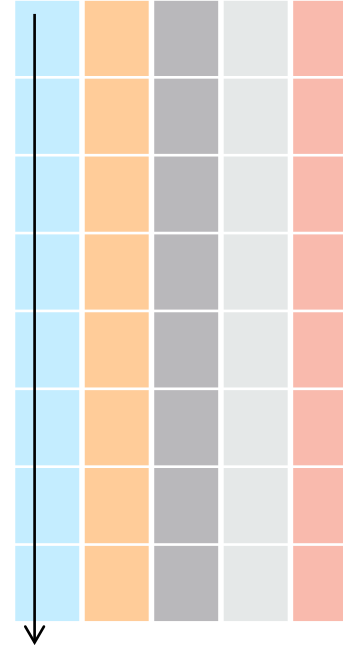
Row vs Columnar Storage

Row Storage



- Traditional Database Storage Method
- Requires all data be read on query
- Limited compression possible

Columnar Storage



- HP Vertica Database Storage Method
- Speeds Query Time by Reading Only Necessary Data
- Ready for Compression

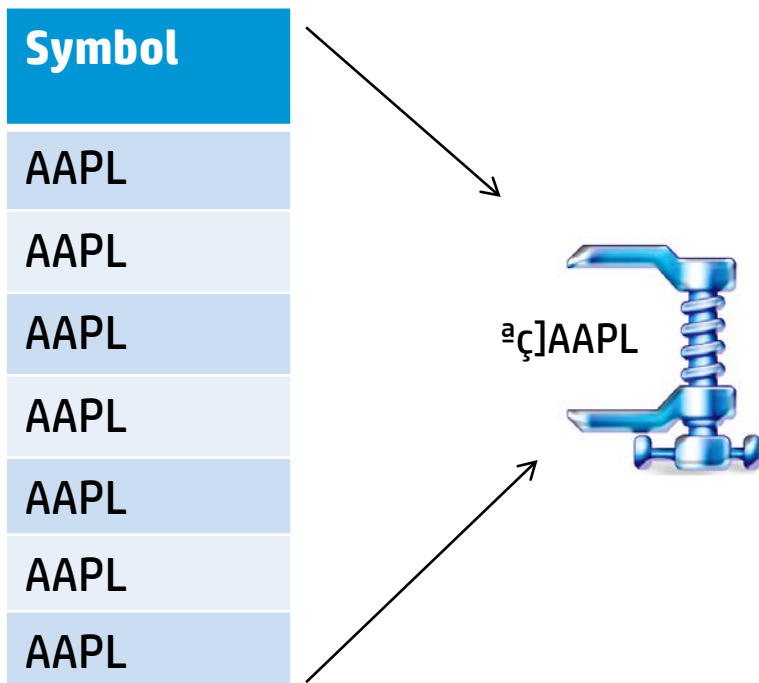
Example

SELECT SUM(volume) FROM trades WHERE symbol = 'HPQ' AND date = '5/12/2014'

Symbol	Date	Time	Price	Volume	Type
AAPL	5/12/2014	13:01:07	530.25	20	Market
AAPL	5/12/2014	13:01:09	530.25	5	Market
GOOG	5/12/2014	13:01:10	1220.71	150	Limit
GOOG	5/12/2014	13:01:11	1220.71	1021	Buy Stop
HPQ	5/12/2014	13:01:11	32.09	50	Limit
HPQ	5/12/2014	13:01:13	32.09	1230	Market
IBM	5/12/2014	13:01:14	187.84	122	Market



Compression



- Compression lowers costly I/O to boost overall performance
- Vertica uses lossless compression
 - integer packing for unencoded integers
 - Lempel–Ziv–Oberhumer for compressible data
- Fast and distributable algorithms – splittable across MPP

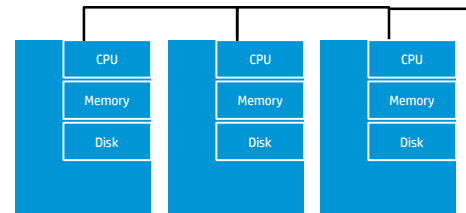
Clustering / MPP / Scale-Out

Parallel design

- Enables distributed storage and workload with active redundancy
- Automatic replication, failover and recovery

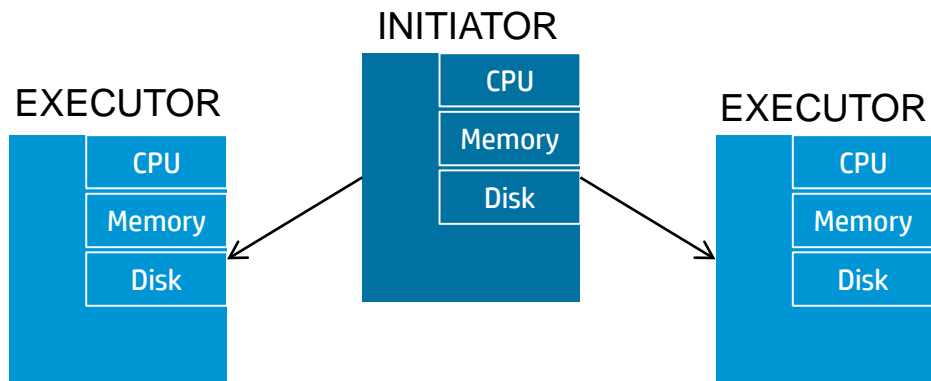
Shared-nothing database architecture

- Provides high scalability on clusters
- No name node or other single point of failure
- Add nodes to achieve optimal capacity and performance
- Lower data center costs, higher density, scale-out



Distributed Query Execution

1. Client connects to a node and issues a query
 - Node the client is connected to becomes the *initiator* node
 - Other nodes in the cluster become *executor nodes*
2. Initiator node parses the query and picks an execution plan
3. Initiator node distributes query plan to executor nodes
4. Initiator node aggregates results from all nodes
5. Initiator node returns final result to the user



Nodes are Peers

- Any node can be the initiator
- No name node or single point of failure
- Query/Load to any node
- Continuous/ real-time load and query

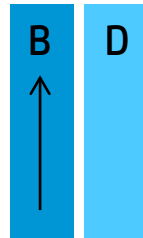
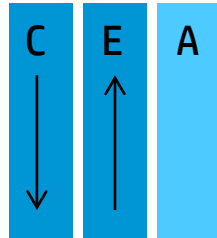
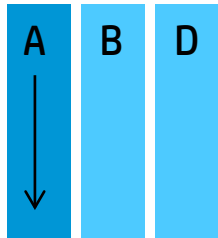
Projections

A	B	C	D	E	...

Projections

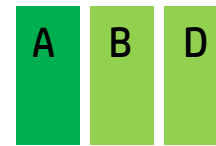
Live Aggregate Projections

**SORT
ORDER**



```
SELECT a,b,d FROM  
projection_name ORDER BY a;
```

**MOST
RECENT**



```
SELECT SUM(A), MIN(B), MAX(D)  
FROM projection_name
```



Query Optimization Comparison

Traditional Materialized Views

- Are secondary storage
- Are rigid: Practically limited to columns and query needs, more columns = more I/O
- Are mostly batch updated
- Provide high data latency

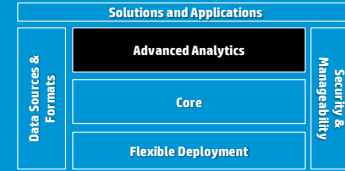
Vertica Projections

- Are primary storage – no base tables are required
- Can be segmented, partitioned, sorted, compressed and encoded to suit your needs
- Have a simple physical design
- Are efficient to load & maintain
- Are versatile – they can support any data model
- Allow you to work with the detailed data
- Provide near-real time low data latency
- Combine high availability with special optimizations for query performance

Traditional Indexes

- Are secondary storage pointing to base table data
- Support one clustered index at most – tough to scale out
- Require complex design choices
- Are expensive to update
- Provide high data latency





Advanced Analytics



Business Impact of Advanced Analytics

We have advanced analytical functions that can speed the development process:

- Event-based Windows
- Monte Carlo
- Time Series Analysis
- Regression Testing
- Geospatial/Place
- Social Media/Pulse
- Much more

What analytical functions mean:

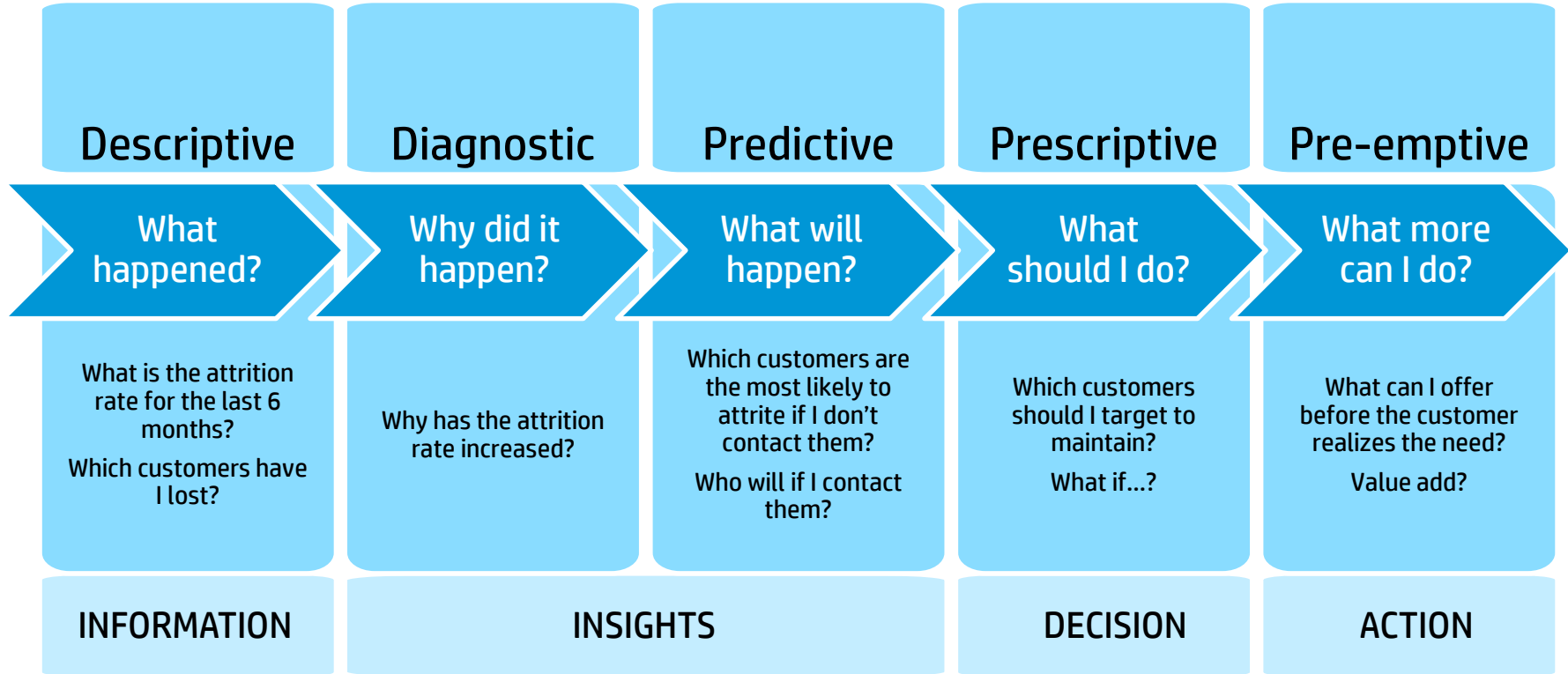
Use to take	Now takes
Hundreds of lines of code	Fewer lines of code

```
67 actions = load '/user/hive/warehouse/actions.db${processingDate}' using PigStorage('uRMM') as (uidId chararray, publisher_u
68 actions_uid = foreach actions generate uidId, uid_removeNull(country) as country, uid_parseDay(time) as day, publisher
69 actions_uid_filtered = filter actions_uid by day == '${processingDate}';
70
71 actions_join_day = JOIN actions_uid_filtered BY day, day_dimension_id BY calendar_date USING 'replicated';
72 actions_join_day -> foreach actions_join_day generate uidId, country, day_key, app_id, source, paid_conversion;
73
74 actions_join_app = JOIN actions_join_day BY app_id, app_dimension_id BY app_id USING 'replicated';
75 actions_join_app -> foreach actions_join_app generate uidId, country, day_key, app_key, platform_key, source, paid_conversion;
76
77 actions_join_country_1 = JOIN actions_join_app BY country LEFT, country_1 BY absolute_country_code USING 'replicated';
78 actions_join_country_2 = JOIN actions_join_country_1 BY country LEFT, country_dimension_id BY country_iso3663_code USING 'r
79
80 converter = foreach actions_join_country_2 generate source as source, uidId as uidId, day_key as day_key, app_key as app_key,
81
82
83
84
85
86
87
88
89
90
91
92
93
94
95
96
97
98
99
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134
135
136
137
138
139
140
141
142
143
144
145
146
147
148
149
150
151
152
153
154
155
156
157
158
159
160
161
162
163
164
165
166
167
168
169
170
171
172
173
174
175
176
177
178
179
180
181
182
183
184
185
186
187
188
189
190
191
192
193
194
195
196
197
198
199
200
201
202
203
204
205
206
207
208
209
210
211
212
213
214
215
216
217
218
219
220
221
222
223
224
225
226
227
228
229
230
231
232
233
234
235
236
237
238
239
240
241
242
243
244
245
246
247
248
249
250
251
252
253
254
255
256
257
258
259
260
261
262
263
264
265
266
267
268
269
270
271
272
273
274
275
276
277
278
279
280
281
282
283
284
285
286
287
288
289
290
291
292
293
294
295
296
297
298
299
300
301
302
303
304
305
306
307
308
309
310
311
312
313
314
315
316
317
318
319
320
321
322
323
324
325
326
327
328
329
330
331
332
333
334
335
336
337
338
339
340
341
342
343
344
345
346
347
348
349
350
351
352
353
354
355
356
357
358
359
360
361
362
363
364
365
366
367
368
369
370
371
372
373
374
375
376
377
378
379
380
381
382
383
384
385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405
406
407
408
409
410
411
412
413
414
415
416
417
418
419
420
421
422
423
424
425
426
427
428
429
430
431
432
433
434
435
436
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455
456
457
458
459
460
461
462
463
464
465
466
467
468
469
470
471
472
473
474
475
476
477
478
479
480
481
482
483
484
485
486
487
488
489
490
491
492
493
494
495
496
497
498
499
500
501
502
503
504
505
506
507
508
509
510
511
512
513
514
515
516
517
518
519
520
521
522
523
524
525
526
527
528
529
530
531
532
533
534
535
536
537
538
539
540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
585
586
587
588
589
590
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629
630
631
632
633
634
635
636
637
638
639
640
641
642
643
644
645
646
647
648
649
650
651
652
653
654
655
656
657
658
659
660
661
662
663
664
665
666
667
668
669
670
671
672
673
674
675
676
677
678
679
680
681
682
683
684
685
686
687
688
689
690
691
692
693
694
695
696
697
698
699
700
701
702
703
704
705
706
707
708
709
710
711
712
713
714
715
716
717
718
719
720
721
722
723
724
725
726
727
728
729
730
731
732
733
734
735
736
737
738
739
740
741
742
743
744
745
746
747
748
749
750
751
752
753
754
755
756
757
758
759
760
761
762
763
764
765
766
767
768
769
770
771
772
773
774
775
776
777
778
779
780
781
782
783
784
785
786
787
788
789
790
791
792
793
794
795
796
797
798
799
800
801
802
803
804
805
806
807
808
809
810
811
812
813
814
815
816
817
818
819
820
821
822
823
824
825
826
827
828
829
830
831
832
833
834
835
836
837
838
839
840
841
842
843
844
845
846
847
848
849
850
851
852
853
854
855
856
857
858
859
860
861
862
863
864
865
866
867
868
869
870
871
872
873
874
875
876
877
878
879
880
881
882
883
884
885
886
887
888
889
890
891
892
893
894
895
896
897
898
899
900
901
902
903
904
905
906
907
908
909
910
911
912
913
914
915
916
917
918
919
920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938
939
940
941
942
943
944
945
946
947
948
949
950
951
952
953
954
955
956
957
958
959
960
961
962
963
964
965
966
967
968
969
970
971
972
973
974
975
976
977
978
979
980
981
982
983
984
985
986
987
988
989
990
991
992
993
994
995
996
997
998
999
1000
```

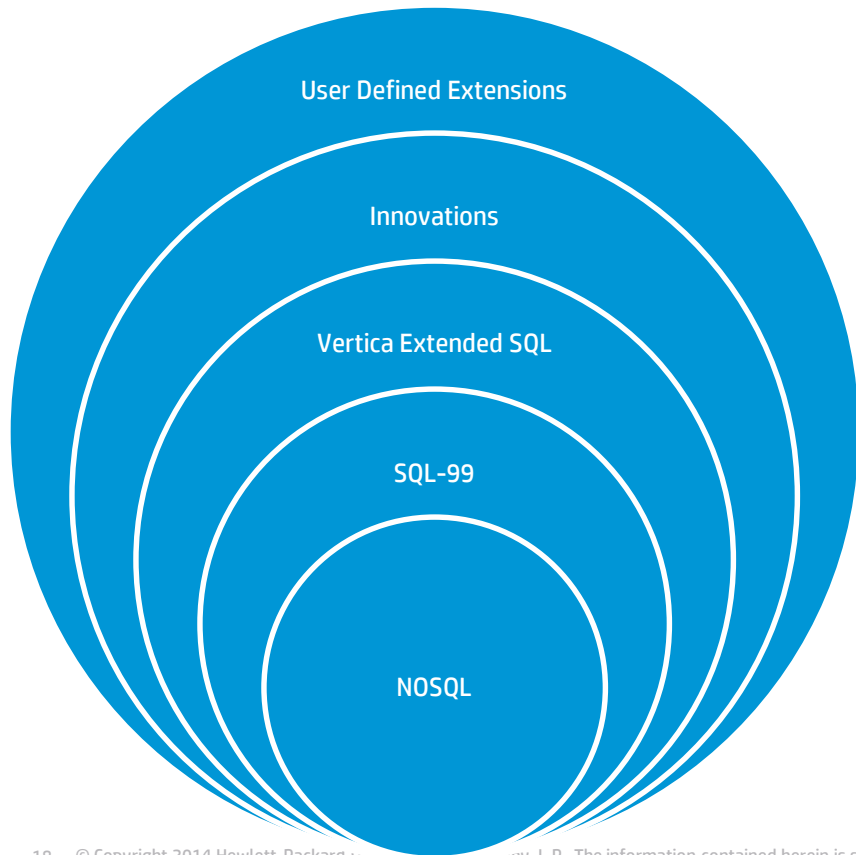
```
SELECT CustomerName.City FROM Customers;
INSERT INTO Customers (CustomerName, ContactName, Address, City, PostalCode, Country) VALUES
('Cardinal', Tom B. Erichsen, 'Skagen 21', 'Cambridge', '02140', USA);
SELECT CustomerName.State FROM Customers;
INSERT INTO Customers (CustomerName, ContactName, Address, City, PostalCode, Country) VALUES
('Cardinal', 'Massachusetts');
```



From Hindsight to Insight to Foresight



Know SQL before you NoSQL



- Not all versions of SQL are the same
- Analytic Functions are limited in:
 - NoSQL implementations
 - Low-end databases
- Vertica provides most complete SQL
 - On RDBMS
 - On Hadoop

Shortcomings of Many Big Data Analytical Solutions

Item	Description	Example Use Case
WHERE clause subqueries	Fast calculations on data that you retrieve using SQL Windowing functions	<ul style="list-style-type: none">• Group clickstream data into sessions to analyze a web visitor's browsing behavior.• Calculate moving-average for stock ticker data.
Data Types	Defining attributes with data types, such as money, time, and JSON	<ul style="list-style-type: none">• Validating Date and Time Values• Calculating Time Difference• Understanding which monetary units were used for a transaction• Understanding partial monetary units, like cents, from whole monetary units. For example, 3.15 specifies 3 dollars and 15 cents.
MERGE-JOIN	Joining data from separate tables	<ul style="list-style-type: none">• Looking up vendor information from a vendor table – ERP and CRM mash up• Mashing up CRM data with sentiment analysis• Leveraging reference data• Many, many uses
Geospatial Functions	Analysis of data with understanding of geometry and/or geography	<ul style="list-style-type: none">• Understanding if an object's location is inside or outside a zone/circle/area• Latitude and longitude applications
Sentiment Scoring Functions	Parsing social media data to understand its relationship to negative/positive words and phrases	<ul style="list-style-type: none">• Acquiring data from twitter and performing sentiment analysis



Analytical Features of Vertica

Vertica SQL Standard SQL-99 Conventions	Vertica Extended-SQL Advanced Analytics with SQL	Vertica Innovations Advanced Analytics using Custom Logic	Vertica User Defined Extensions
Aggregate	Sessionization	Regression Testing	Analytics <ul style="list-style-type: none"> • C++ • Java • R Connection <ul style="list-style-type: none"> • ODBC/JDBC • HIVE • Hadoop • Flex Zone
Analytical	Time Series <ul style="list-style-type: none"> • Time slice • Interpolation (Constant & Linear) • Gap Filling • Aggregate 	Statistical Modeling	
Window Functions	Event-based Windows <ul style="list-style-type: none"> • Conditional Change Event • Conditional True Event 	Classification Algorithms	
Graph	Event Series Joins	Page Rank	
Monte Carlo	Social Media/Pulse <ul style="list-style-type: none"> • Text Mining • Patterns/Trends 	Text-mining	
Geospatial	Pattern Matching <ul style="list-style-type: none"> • Match, Define, Pattern Keywords • Funnel Analysis 	Geospatial (Place)	
Statistical			



Solutions and Applications

Solutions and Applications		
Data Sources & Formats	Advanced Analytics	Security & Manageability
	Core	
	Flexible Deployment	



HP Vertica Pulse

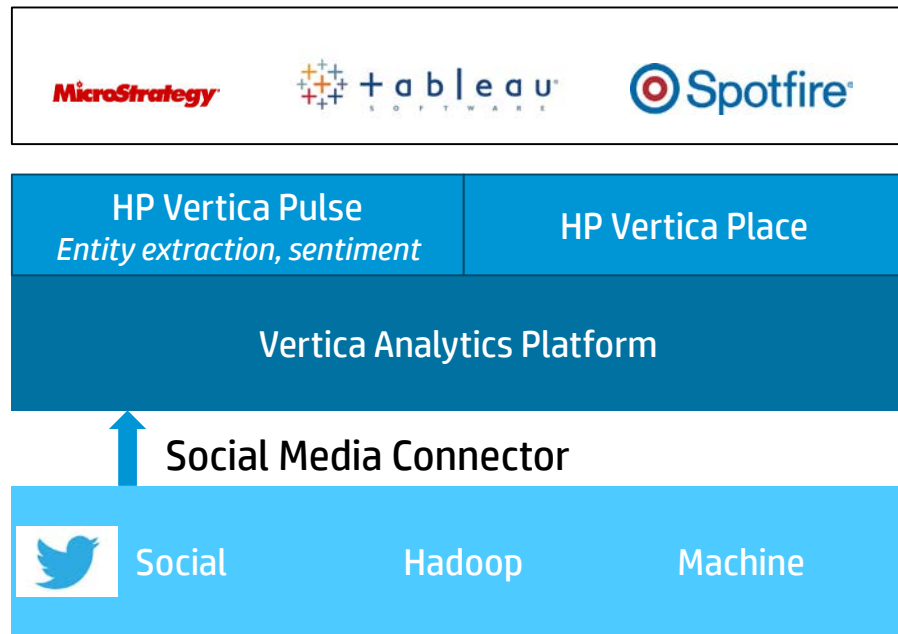
Perform Sentiment Analysis

Challenge: Performing sentiment analysis is time-consuming and tedious

SOLUTION: HP Vertica Pulse

Benefit:

- Scalable, in-database entity extraction and sentiment analysis
- Aggregate and drill-down views
- Easy to get started



HP Vertica Distributed R

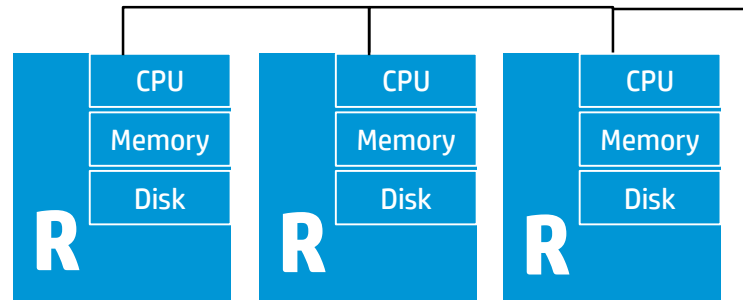
R-based Analytics

Challenge: Customers want to use R for analytics. However, R scalability is always a question

SOLUTION: HP Distributed R

Benefit:

- Analyze data sets too large for standard R
- Perform complex analyses much more quickly (20x faster than Hadoop)
- Use familiar R environment to explore data, develop, and execute algorithms
- Operate on full data set (no down sampling)



Algorithm	Use cases
Linear Regression (GLM)	Risk Analysis, Trend Analysis, etc.
Logistic Regression (GLM)	Customer Response modeling, Healthcare analytics (Disease analysis)
Random Forest	Customer churn, Market campaign analysis
K-Means Clustering	Customer segmentation, Fraud detection, Anomaly detection
Page Rank	Identify influencers



Vertica Place

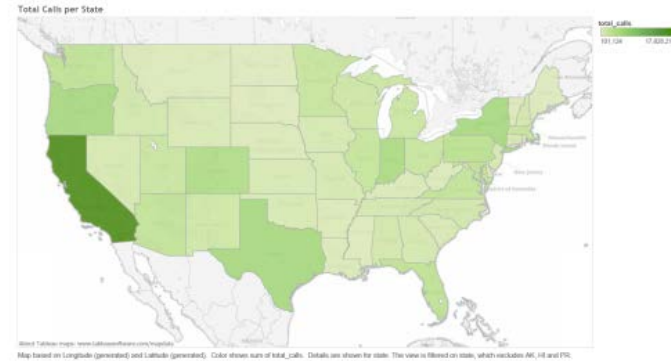
Geospatial Analysis

Challenge: Analysis of data with understanding of geometry and/or geography

SOLUTION: HP Vertica Place

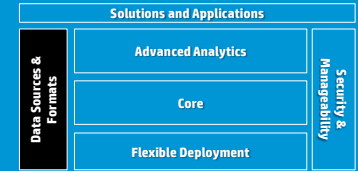
Benefit:

- Optimized Spatial Joins with memory-resident Geospatial Indexing replaces expensive scans with easy look-ups
- Easy to use OGC-standard based implementation with spatial functions to compute: distance, intersections
- Simple integration with third party applications
 - ESRI Shapefile parser
 - Internal functions to ingest WKT/WKB data



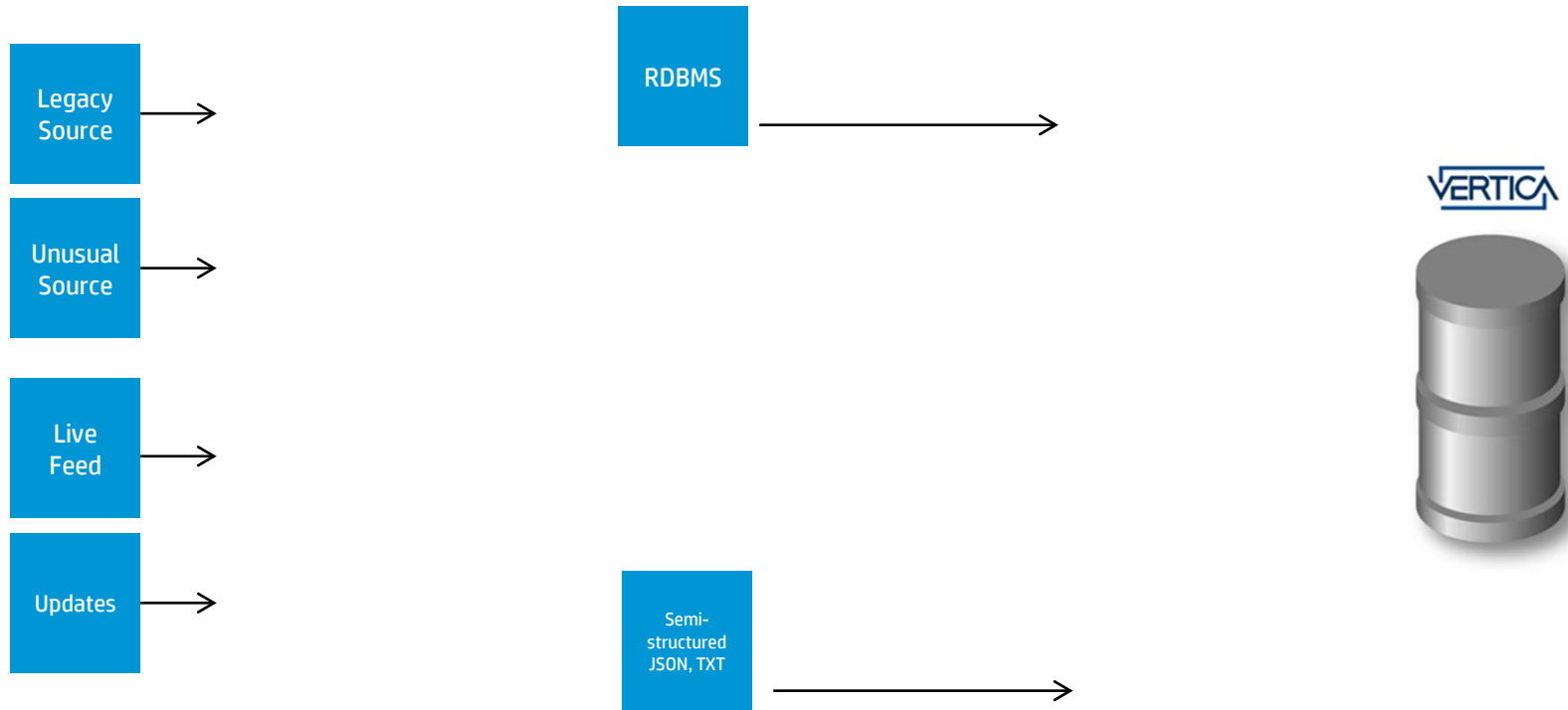
```
SELECT STV_Intersect(gid, geom
    USING PARAMETERS
        index='/dat/states.idx')
OVER()
AS (call_gid, state_gid)
FROM calls;
```



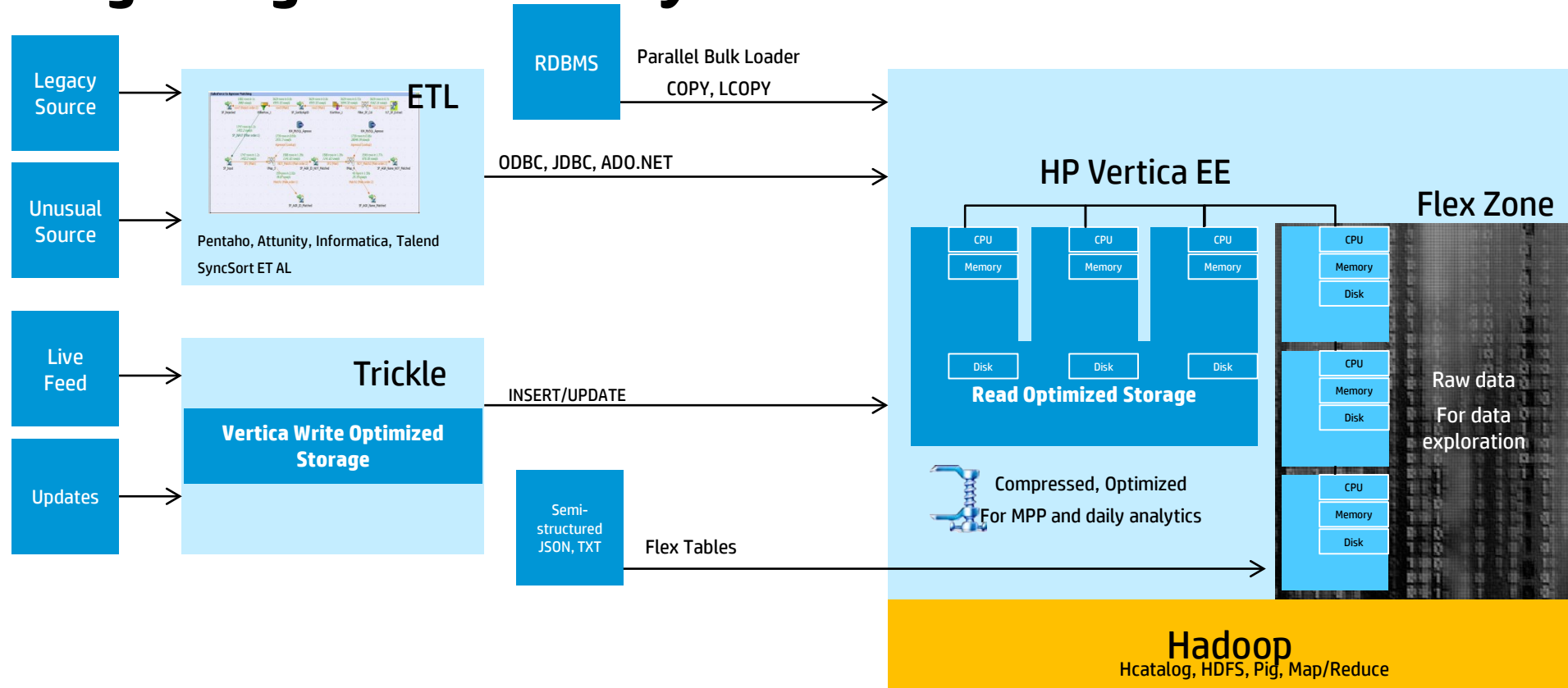


Data Ingestion and Storage

Ingesting Data for Analytics



Ingesting Data for Analytics



Data Storage Transparency

Query Data in your Hadoop Data Lake or Optimized HP Vertica Storage



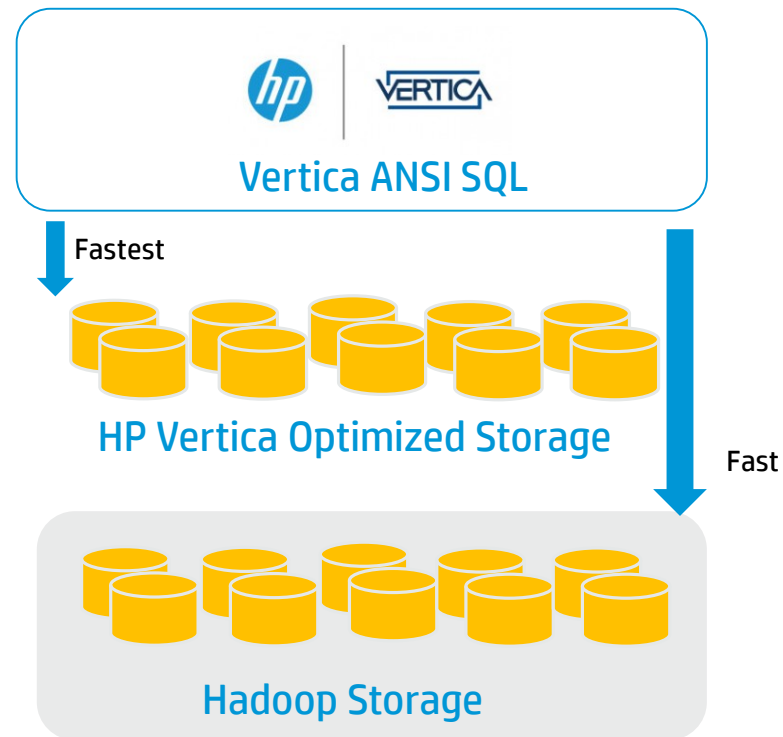
Analyst shouldn't need to care where the data is located, or how it is formatted



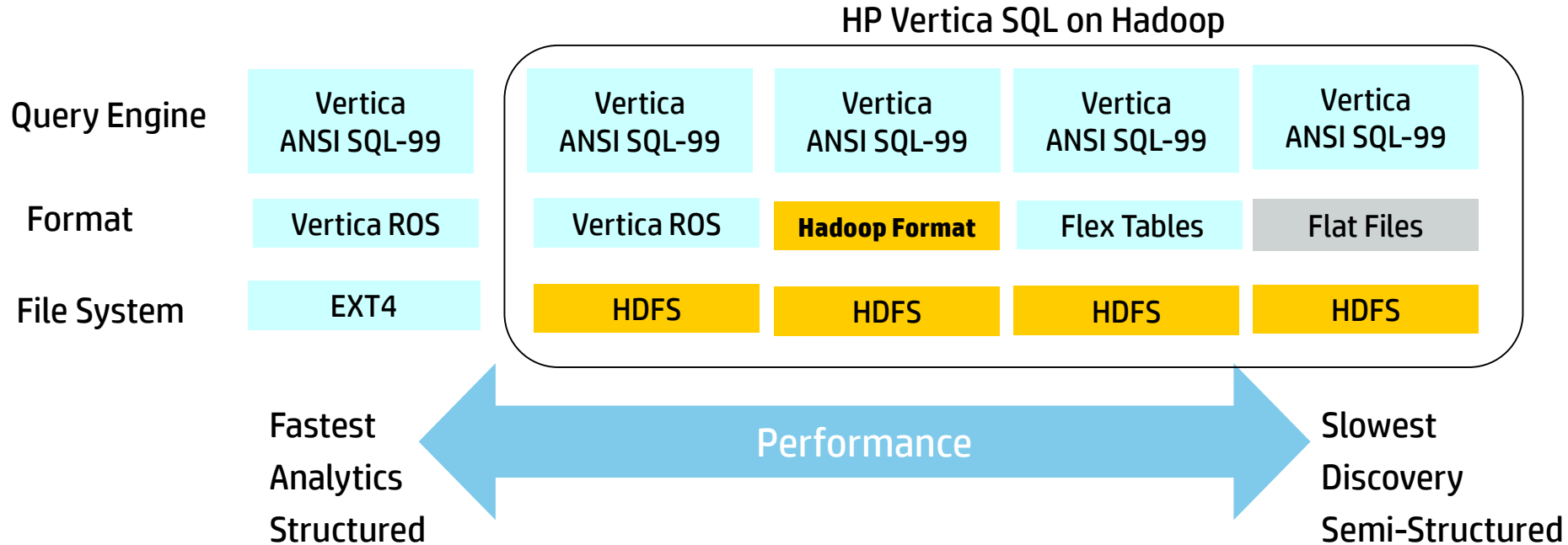
DBA's can use multiple storage pools for convenience and to match cost and performance needs



The data engineer has multiple options to ingest and/or move data between Vertica and Hadoop to suite cost and performance needs



Data Storage Options and Performance



Hadoop Agnostic

Powerful SQL on Hadoop for any Hadoop Distribution

Works with:

- Cloudera Distribution Including Apache Hadoop (CDH) 4.6, 5.0, and 5.0.1
- Hortonworks Data Platform (HDP) 2.1
- MapR 3.1.1 and 3.0.3

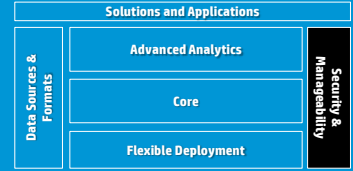
Also:

- MapR NFS may be used as storage

Versions update frequently. Check online documentation for a detailed description and updates to compatibility information.

cloudera



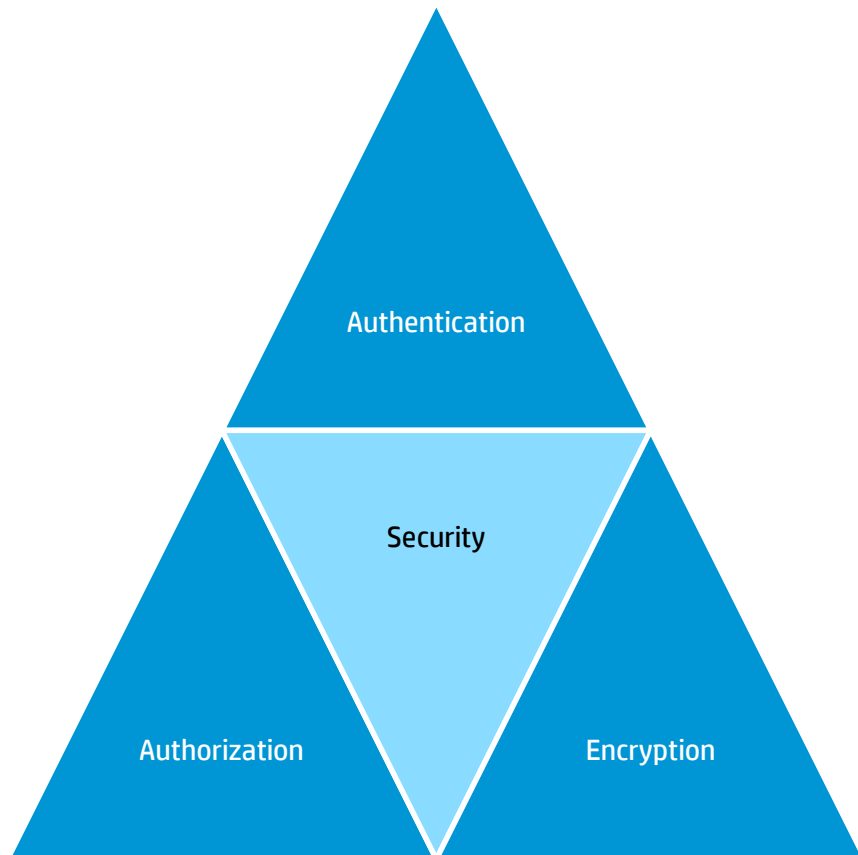
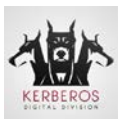


Security & Manageability



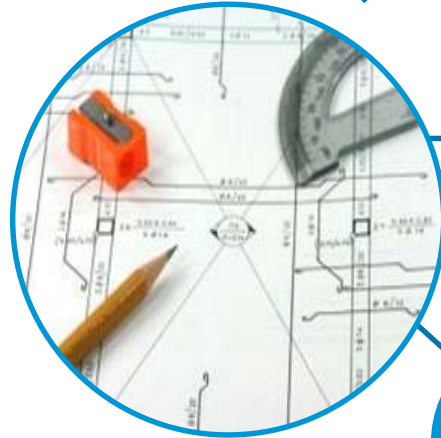
Security

- SSL security protocol to authenticate, encrypt and verify access
- Use SQL to GRANT/REVOKE to database, schema, table, views, insert/update/delete, etc.
- Manage Users in the Vertica Management Console or use LDAP/Kerberos for external management



Database Designer

Database Designer



Leverage

- Typical Queries
- Sample Data/Schemas
- Historical Statistics and Logs

Optimize

- Query Performance
- Data Loading
- Storage Footprint

Benefit

- Faster Queries
- Lower Hardware Costs
- Shortened design time
- Lower costs to maintain and optimize

Workload Management

Problem: Rogue queries can take over

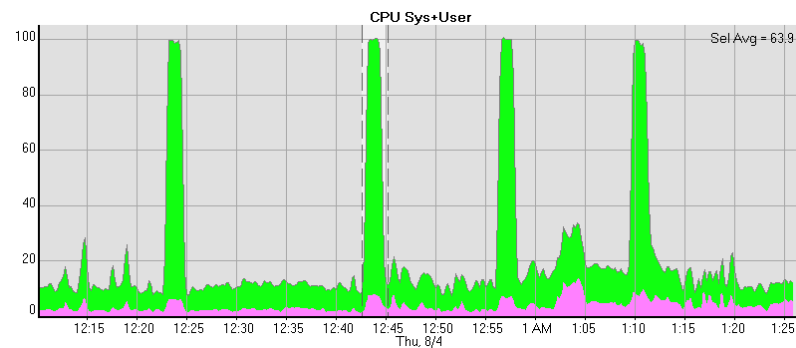
- Affect loads or tactical queries
- Steal Resources from Daily Analytics
- Disrupt Business Processes

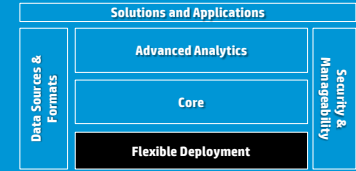
Solution: Workload Management

- Reserve resources for high-priority queries
- Apply run-time prioritization to manage CPU and I/O
- Take control of rogue queries



```
SELECT "Physical Exam".ParticipantId,"Physical Exam".SequenceNum,"Physical Exam".Date,"Physical Exam".Day,"Physical Exam".Weight_kg,"Physical Exam".Temp_C,"Physical Exam".SystolicBloodPressure,"Physical Exam".DiastolicBloodPressure,"Physical Exam".Pulse,"Physical Exam".Respirations,"Physical Exam".Signature,"Physical Exam".Pregnancy,"Physical Exam".Language,AverageTempPerParticipant.AverageTemp, FROM "Physical Exam" INNER JOIN AverageTempPerParticipant ON "Physical Exam".ParticipantID=AverageTempPerParticipant.ParticipantID
```





Flexible Deployment



HP Vertica makes data matter

Purpose built for Big Data from the first line of code



Real time analytics

Gain insight into your data
50x-1,000x faster than legacy
products



Massive scalability

Infinitely scale your solution
by adding an unlimited
number of low cost nodes



Open architecture

Built-in support for Hadoop, R,
and a range of ETL and BI tools



Optimized data storage

Store 10x-30x more data per
server than row databases
with patented columnar
compression



Software Only



Private Cloud



Public Cloud



Appliance

