

Perth MS Cloud User Group  
December 2016

# Azure SQL DW

## Lessons from the Field

Bhavik Merchant  
Data Platform Solution Architect  
Microsoft Australia  
[bhmerc@microsoft.com](mailto:bhmerc@microsoft.com)



# Session Goals

- Drink beer and eat pizza ;)
- Know what SQL DW is
- Understand real client scenario
  - Focus on getting data in
- Follow my journey and lessons learnt
- Get some guidance to take into practice
- Learn about some limits

# Overview



# What is Azure SQL DW?

- Cloud based, scale out DB
- Designed for **massive** data volumes
  - Approx 5x compression. Support 240 Tb **compressed** user data
- Built on MPP architecture
- Separates compute and storage
- Flexible: Scale out, scale back, pause/resume
- Allows seamless querying over Hadoop
  - Can leverage this with minimal setup

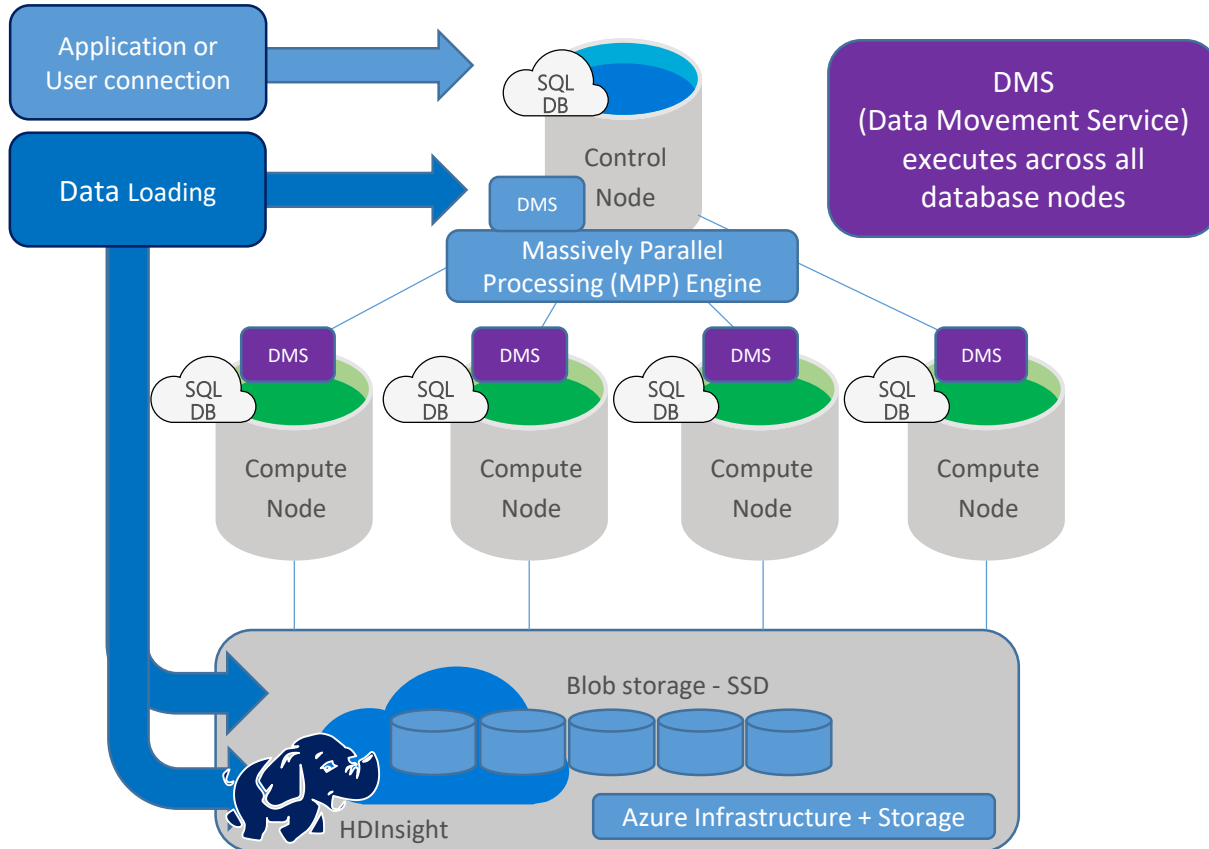
# What are the Benefits?

- Deploy within 10 minutes
  - Almost zero architecting, configuring, tuning\*
- Use familiar paradigms
  - SQL tables, stored procs, indexes, partitions
  - T-SQL
- Cost control via elasticity – think peak/off-peak
- Hybrid architectures possible (sensitive vs non)
- Familiar analytical tools – Power BI, Excel, SSRS

# Internals



# What's in the Box?



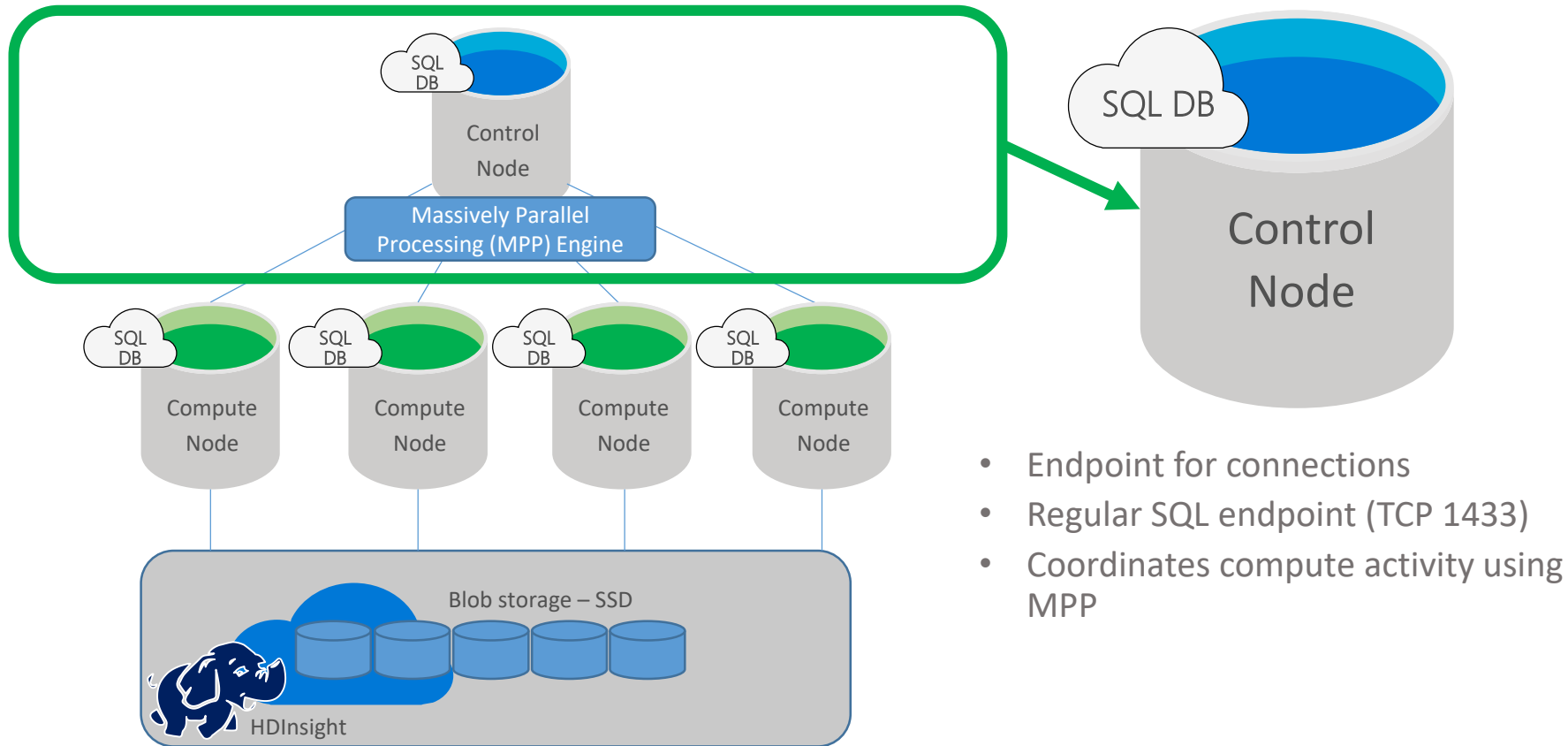
Storage and Compute are de-coupled,  
enabling a true elastic service and  
separate charging for both compute  
and storage

Compute  
Scale compute up or down  
when required

Pause, Restart, Stop, Start.

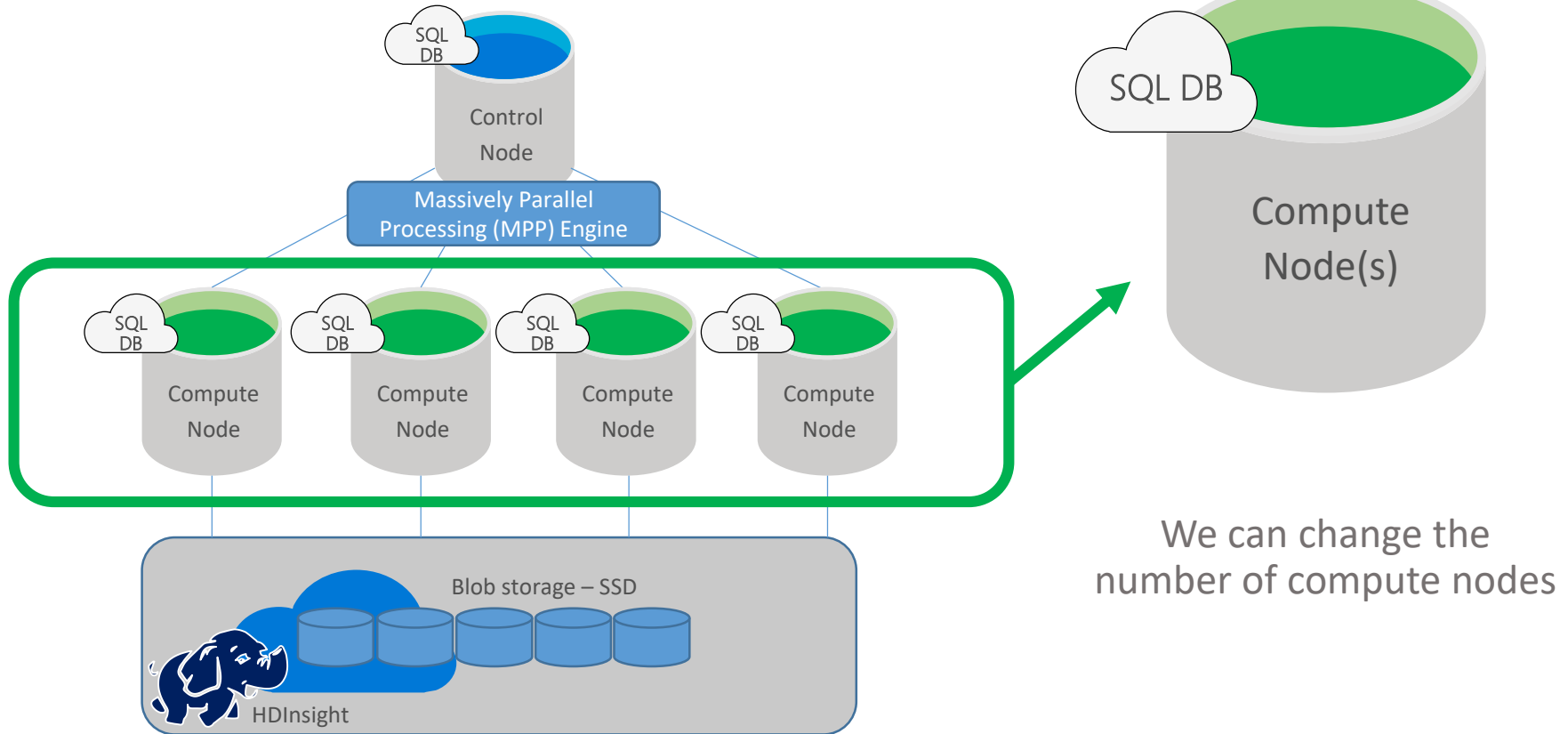
Storage  
Add/Load data to WASB(S)  
without incurring compute  
costs

# Control Node

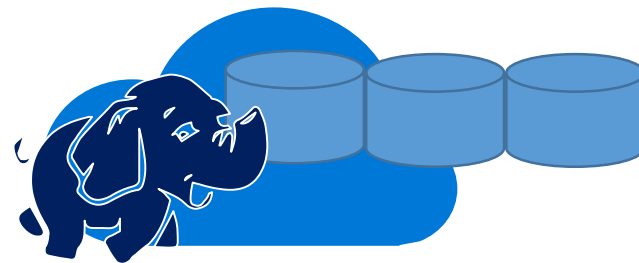
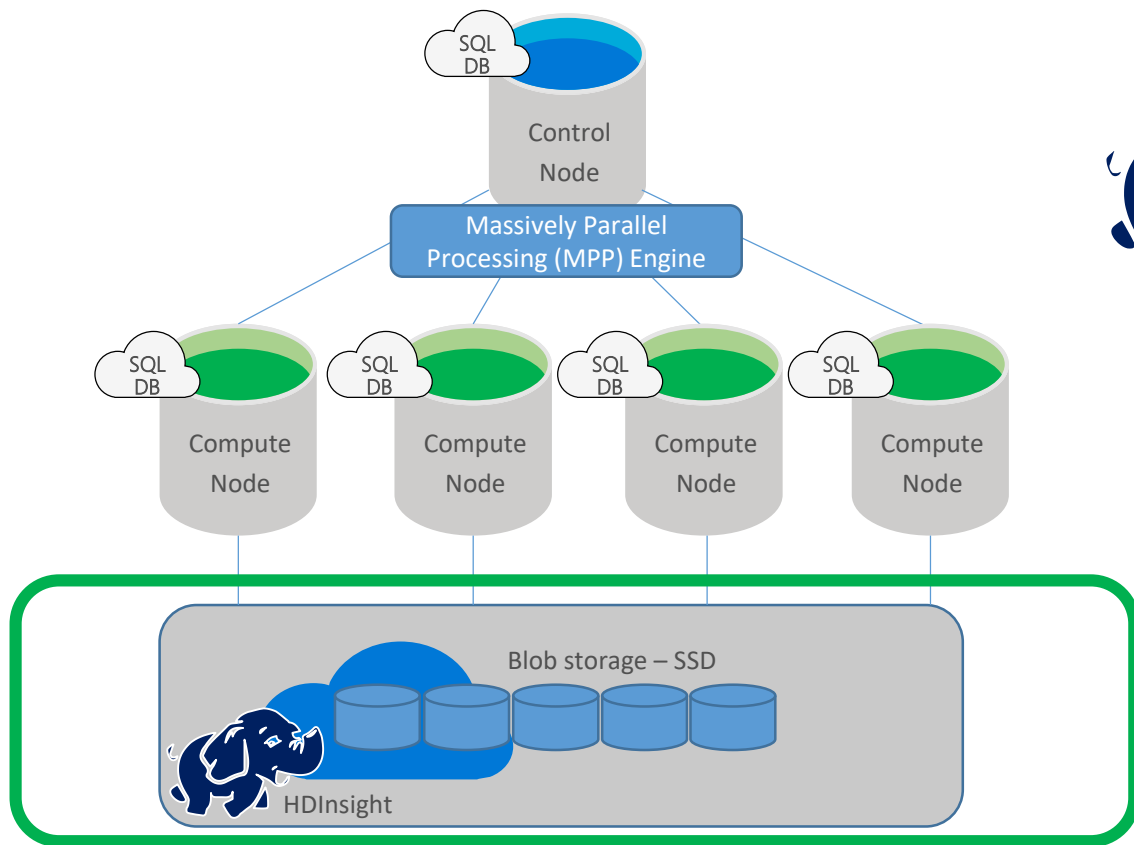




# Compute Nodes



# Blob storage



- RA-GRS storage
- +PB's of storage
- Load data without incurring compute costs

# Customer Scenario



# Business Problem

- Anomaly detection on gas plant sensors
- Platform must cater for a range of questions
- Limited self-service capability
- You don't know what you don't know
- Queries are ad-hoc and varied - mainly around
  - I want to look at one tag over a range of day/weeks/months
  - I want to look at a group of tags over smaller period
  - I want to be able to average a range of tag values in the same category
  - Etc

# Technical Challenges

- We have billions of rows of sensor data
- How to load data to cloud efficiently?
- How to design for performance for wide range of questions?
- How to visualise massive datasets?
  - Can we cater for the inadvertent troublesome user?
  - Even on 4K screen, you can visualise 10s of thousands of points in a trend

# My Constraints

- Use existing data formats
- Load the data in a very short time
- Solve the process and visualisation challenges
  - Beyond the scope of this presentation
- The focus of the presentation is getting the data in

# Loading Data



# Source Data

- Asked the client to dump data to Azure Blob storage
- Millions of gzip files
- Each zip contained ~200-1500 PSV files
- Tens of thousands of “folders”



# A Look at the Data

1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 33 34 35 36 37 38 39 40 41 42 43 44 45 46 47 48 49 50 51 52 53 54 55 56 57 58 59 60 61 62 63 64 65 66 67 68 69 70 71 72 73 74 75 76 77 78 79 80 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 1001 1002 1003 1004 1005 1006 1007 1008 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 104

# Polybase to the Rescue? In theory...

- Configure credentials and external data source in SQL DW
- Configure external file format in SQL DW
- Define external table with file format
- Load from Blob to SQL DW using CTAS
- Once in DW, get columnstore compression and usual T-SQL benefits

# In reality...

- Polybase relies on Hadoop
- Creating external data source killed SQL DW
- Discovered Hadoop limitation
  - 33k files/folders or files per folder

# LESSON 1

- Do not overload Polybase/Hadoop with over 33k folders/files per folder



# Loading Data... Again



# Need a Better Distribution Strategy

- How do I spread the millions of gzip files?
- Look for patterns
- Move more files into less folders
- Try load again

# The New Plan

- PowerShell to dump list of files from Blob store
- Bulk load file list to SQL table (SSIS via Wizard)
- Find groups and make a subgroup (pattern) list for each group (T-SQL query)
- Dump patterns to CSV files (SQLCMD via T-SQL)
- Read each pattern in group and execute AZCopy (in parallel via PowerShell)

# Visualising the patterns

- 20 Groups -> 100-200 batches -> hundreds of files

Group	Batch (lines in CSV)	Example Filenames that AZCopy would match
<b>EVL01Value.csv</b>	<b>EVL01Value</b> 1380606	<b>E01Value13080606</b> 7603720000.xml_17130.psv.gz <b>E01Value13080606</b> 1972930000.xml_72310.psv.gz <b>E01Value13080066</b> 7603720000.xml_13049.psv.gz
	<b>EVL01Value</b> 1380708	<b>E01Value13080708</b> 7603720000.xml_17130.psv.gz <b>E01Value13080708</b> 3849810000.xml_95829.psv.gz <b>E01Value13080708</b> 8422340000.xml_24524.psv.gz
<b>EVL02Value.csv</b>	<b>EVL02Value</b> 1380606	<b>E02Value13080606</b> 7603720000.xml_17145.psv.gz <b>E02Value13080606</b> 1972930000.xml_84836.psv.gz <b>E02Value13080066</b> 7603720000.xml_23423.psv.gz
	<b>EVL02Value</b> 1380708	<b>E02Value13080708</b> 7603720000.xml_66239.psv.gz <b>E02Value13080708</b> 3849820000.xml_16455.psv.gz <b>E02Value13080708</b> 8422340000.xml_17653.psv.gz



## Lets see the Code

[illegible]

# More Problems...

- AZCopy is multi-threaded and parallel
  - Cleverly generates storage API calls for you
- BUT
  - It takes forever to do each group!
  - I have limited time
  - I have limited patience
- Can I use faster storage?
- Why not run the groups in parallel?

# Modified Approach

- Move to Premium storage (SSD) to make life easier later
- I have 20 groups so..
- Run each group in parallel
  - Do it on an Azure VM
  - Destinations are new premium storage accounts (helps later)
  - 20 CMD windows running PowerShell AZCopy loop to Premium storage
- Sounds good now... right?

# Yet More Problems

- PROBLEM 1: Premium storage = only Page Blob
  - AZCopy cant change from Block blob to Page blob on the fly
- PROBLEM 2: "Existing manifest", "incomplete operation" error in AZCopy
  - AZCopy cant be run in parallel under the same user profile
- Clever parallelisation no more 😞

# LESSON 2

- Plan storage account tiers and number beforehand!



# LESSON 3

- AZCopy is parallel by default, but cant be run in parallel itself!



# Loading Data...

## Still Going



# Divide and Conquer Again!

- Create 10 windows users on Azure VM
- Run a group under each user in parallel
  - Separation of profile means no manifest error
- Wake up in the morning and kick off the other 10
- Maxxed out a G Series VM with 32 cores for about 10 hours



# Loading Data...

## Last Step



# Finally, Load to SQL DW!

- Set up external location again
- Use CTAS to stage the data

```
CREATE TABLE [staging].[table_name]  
WITH (DISTRIBUTION = ...)  
AS  
SELECT ... FROM external_table
```

# One more Hiccup!

- Initial table load was pretty fast (CTAS)
- Subsequent loads used

```
INSERT INTO [staging].[table_name]  
SELECT ... FROM external_table
```

# LESSON 4

- CTAS is minimally logged and fast. Use it wherever possible. Subsequent inserts are slower.



# Where Does SQL DW Fit?

SQL Server VM  
(IaaS)

Azure SQL Database

Azure SQL Data  
Warehouse

Azure Data Lake

OLTP / DW workloads  
Lift and Shift  
Customer managed  
Shared features

1TB+

OLTP/ DW workloads  
Net new development  
Fully managed service  
No shared features

1Gb-1Tb GB

DW workloads only  
Fully managed  
Dynamic Pause/Scale

250GB – PB+

Non-relational  
Cheap, flexible Access  
Processing raw data

1 TB+

# Some SQL DW Caveats

- No PK/FK relationships
  - A result of MPP architecture. Workaround – views
- Other unsupported constructs
  - E.g. Constraints, triggers
- Must manually create stats on columns
  - Don't skip! Choose join/group columns. Regularly update.
- Cant fully use SSMS/SSDT fully at present
- Not all data types supported
  - Guide available online for conversion
- 32k buffer size for Polybase (will be increased)

# Final Takeaways

- Have an initial architectural plan
- Consider each step of the data pipeline
  - Initial load could be very different from incremental updates
- Research limits/restrictions on ALL services
- Azure was very agile
  - Killed SQL DW and provisioned a new one
  - Provisioned additional storage as needed
  - Tested load on D series, scaled up VM to G Series when needed
- Got through it all in 2-3 days

