

Introduction to the Cluster Metadata Store (CMS)

Last updated by | Subbu Kandhaswamy | Sep 17, 2020 at 6:27 PM PDT

Contents

- [Cluster Metadata Store \(CMS\)](#)
 - [Overview](#)
 - [CMS Features](#)
 - [Availability](#)
 - [Domain Model](#)
 - [CMS v1 Schema](#)
 - [State Machine Data](#)
 - [Customer Visible Data](#)
 - [SQL Cluster Infrastructure](#)
 - [Infrastructure](#)

Cluster Metadata Store (CMS)

Overview

The Cluster Metadata Store (CMS) provides a single logical point of metadata storage for a SAWA v2 cluster. Located on the Cluster Control Manager node of the control ring, the store provides the source of 'truth' regarding the state of the cluster and its resources. Each resource being managed in the cluster is represented by a row in a table in CMS that reflects the state of that resource. CMS provides the persistence for state-machine driven workflows driven from within the Management Service that control all updates to resources and which ensures that the metadata either reflects a known stable state of a resource or that a workflow is in-progress currently modifying the resource.

Each type of resource that requires management is represented by a table in the store with each instance of the resource represented by a row in that table. Metadata describing the resource is held in columns on the table. Lower level resources that are managed by Windows Fabric exclusively, such as the location and state of replicas is not represented in the store. Some resources are purely logical. For example, Logical Server is an important user abstraction that hides the location of a database on an instance, allowing databases to be distributed across multiple instances but presented to the user as co-located for management purposes.

As a governing principle of the SAWA v2 architecture, all creation, update and deletion of resources and their associated metadata is mediated by managed workflows that are tightly integrated with the metadata store. Fine-grained state machines are defined for each resource type and are instantiated at runtime to manage workflows impacting those resources. State machines describe stable states and unstable or work-in-progress states. State-related information for each resource is also persisted to the resource table.

At runtime a running state machine is instantiated for each resource being operated on based on the data held in the metadata store. This state machine instance mediates creation, all updates and deletion of the resource. All updates to resources and metadata in the store are handled by actions. Actions can be triggered by an event or automatically on entering into an unstable state. An action may transition the resource into a new state based on its outcome. An action on one resource may raise events on other resources allowing workflows to cascade across multiple related resources. The use of unstable states, automatic actions and branching transitions allows complex workflows to be expressed and managed at runtime. Workflows are triggered by an event on a resource in a stable state and continue until all resources affected by the workflow are in a stable state.

Through the use of unstable states and by persisting every state change to the store, a workflow can be designed to retry if resources are temporarily unavailable, and can be restarted and continued in the event of a failure of one or more executing state machines. Workflows can be designed to detect and automatically recover from errors, if necessary, reversing partial changes to resources.

CMS is the source of truth for the state of all SQL-related resources being managed on a cluster, however CMS is not a cache.

CMS Features

- Source of truth for the state of all SQL-Related resources being managed in the cluster.
- Not a cache. The control ring architecture and connectivity model do not require the metadata store to be involved in the user login pipeline or subsequent SQL execution at runtime.
- It is neither updated optimistically before resource updates nor after the fact.
- For any individual resource CMS reflects either the current state of the resource or that a workflow is in progress intending to create, update or delete the resource.
- CMS supports the metadata requirements of all internal cluster management services and user-focused login and provisioning services.
- CMS enables state-machine based workflows through persisting state information. By persisting transient state changes in durable storage, CMS allows workflows to be resilient to unavailable resources, and enables workflow restart in the event of a failure of a workflow controller, the CCM node or the CMS database or the control ring as a whole.
- CMS is not required for or referenced during user login to a database; user login must succeed even if the CMS store is unavailable.
- CMS is query-able from off-cluster for management, monitoring and troubleshooting.
- CMS is secured so that only authorized users or components can update or query data. The store enables and where required enforces adherence to rules governing PII or other sensitive data.
- CMS is monitoring to provide insight into the immediate health and performance of the store (distinct from monitoring the resources represented in the store).
- Full telemetry data is available from all update actions and state changes allowing off-cluster troubleshooting, audit and performance analysis.

- CMS interactions are designed in a way that it would not affect scenarios requiring performance goals.

Availability

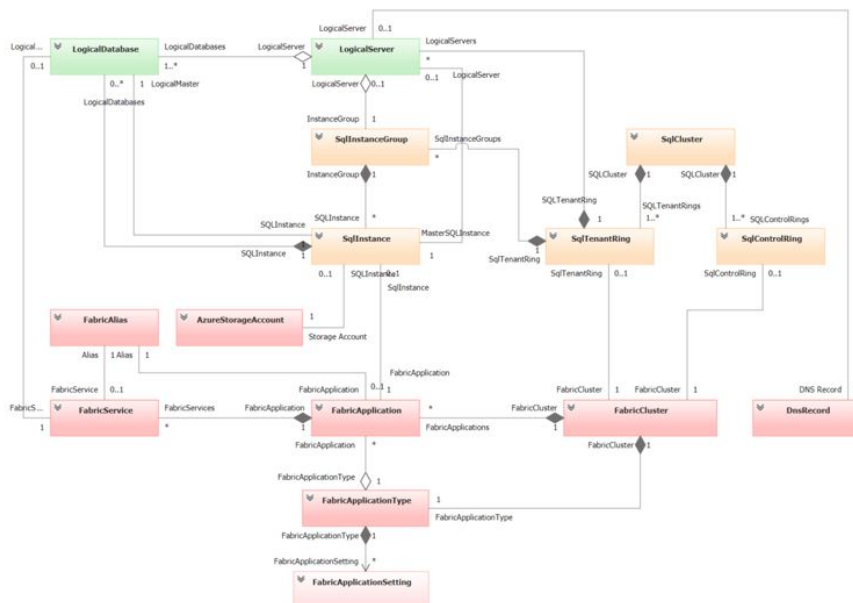
- CMS is a single point of failure in the SAWA v2 architecture. If CMS is unavailable, server and database provisioning and management will be unavailable. It is critical that it achieve maximum reliability.
- As the control ring runs on public compute on SKUs without local database storage, CMS will be implemented on a Remote Storage SQL Database and will be subject to the availability inherent in that storage model.
- As CMS uses Remote Storage, the database files are mastered in Azure Storage and must leverage Azure storage-provided geo-replication.
- CMS is protected from CCM node or Control Ring failure through WinFab.
- CMS is protected from data center disaster and can be recovered from remote backups in the event of failure
- All CMS content must be backed up off-cluster and available for recovery in the event of a cluster failure.

Domain Model

The key concepts that are represented in CMS can usefully be represented in a domain mode. The model can be thought of as organized in layers reflecting the architectural structure of a cluster.



Picture 1



Picture 2

The domain model reflects the structural constraints that apply to a steady-state running cluster after initialization.

CMS v1 Schema

State Machine Data

The following columns are included in all tables managed by a state machine.

Column Name	Data Type	Nullable	Default	Description
state	nvarchar(128)	False		The current state in the FSM state machine.
stable	bit	False		Indicates that the current state is unstable (triggers an automatic action).
error	bit	False		Indicates that the current state is declared as an error state.
request_id	uniqueidentifier	True		<p>The active request id. Populated while the instance is participating in a workflow. This allows correlation of instances of different state machines that are participating in the same request. The value is provided by the client – typically will be provided by RDFE over the RP API.</p> <p>The value is set when an instance joins a workflow, either as the initial state machine or one that joins as a successor through instantiation or an event raised in an action executed on another instance already in the workflow.</p> <p>The value is reset to null when the state machine returns to a stable state.</p>
workflow_position [new]	hierarchyid	True		<p>Hierarchically structured identifier of the position, relative to other state machine instances in the current workflow, at which this instance last received an event or was created.</p> <p>The value is reset to null when the state machine returns to a stable state.</p>

Column Name	Data Type	Nullable	Default	Description
next_successor	int	False	0	<p>The id used for the next successor in a workflow. When an event is raised in an action on another state machine or a state machine is initialized, this value is appended to the current workflow_position to create the workflow position of the successor. The value is then incremented by the FSM framework.</p> <p>The value is reset to 1 when the state machine returns to a stable state.</p>
create_time [new]	datetime2	False		The time the record was created.
last_update_time	datetime2 [was datetime]	False		The time the record was last updated. Has the same value as CreateTime when first created.
last_state_change_time [new]	datetime2	False		The time the state was last changed. May be different from last_update_time. The time is not changed if a transition flows recursively to the same state.
concurrency_token	bigint	True		A value that is changed whenever the row is changed.
process_id [new]		True		The fabric id of the process containing an FSM instance that is currently actively managing the state machine on this row. Is populated when a long running action is in progress that is not-interruptible. Allows a running state machine to be picked up by the same process if the

Column Name	Data Type	Nullable	Default	Description
				process fails and is restarted by WinFab.
last_exception [was exception]	xml	True		Details of the last exception raised during execution of the state machine on this instance. Cleared on next successful action or state transition on the instance.

Customer Visible Data

Logical Databases

Table: tbl_logical_databases

View: logical_databases

FSM: LogicalDatabaseStateMachine

Contains one row per database. Includes Logical Master and User databases.

	Column Name	Data Type	Nullable	Default	Description
PK	logical_server_name [new]	nvarchar(128)	True		The logical server that contains the database. [persisted computed column based on following SQL InstanceGroup -> LogicalServer]
PK	logical_database_id [new]	uniqueidentifier	False	newsequentialid()	Globally-scoped database Id. Allows tracking cross-cluster database moves or relationships, for example.
	sql_instance_name [was InstanceName]	nvarchar(128)	False		Identifies the SQL instance in which the database is hosted. This value may change if the database edition changed between Premium and Standard which would cause the database to be moved between instances. Will also change if the database is moved between instances to load balance instance.
	logical_database_name	nvarchar(128)	False		The name of the database.
	dropped_time	datetime2	True		The time this database was dropped or

	Column Name	Data Type	Nullable	Default	Description
					deferred dropped. This is included in an unique index with logical_server_name, database_name to enforce name uniqueness but allow database names to be reused once dropped.
	edition	nvarchar(25)	False	Standard	Standard Premium
	max_size_bytes	bigint	False	1048576 (1 MB)	Database size quota. If the database exceeds this size it is capped and allows only read and delete queries. Permitted values defined by a set of client side business rules that currently allow 20MB, 100MB, 1GB, 5GB, 10GB, 20GB, 30GB, 40GB, 50GB, 100GB, 150GB. These rules are not enforced on the database.
	service_level_objective	nvarchar(25)	False	Shared	Shared P1 P2 P3 P4 P5 P6 P7 P8 P9 P10 P11 P12 P13 P14 P15 P16 P17 P18 P19 P20 P21 P22 P23 P24 P25 P26 P27 P28 P29 P30 P31 P32 P33 P34 P35 P36 P37 P38 P39 P40 P41 P42 P43 P44 P45 P46 P47 P48 P49 P50 P51 P52 P53 P54 P55 P56 P57 P58 P59 P60 P61 P62 P63 P64 P65 P66 P67 P68 P69 P70 P71 P72 P73 P74 P75 P76 P77 P78 P79 P80 P81 P82 P83 P84 P85 P86 P87 P88 P89 P90 P91 P92 P93 P94 P95 P96 P97 P98 P99 P100 P101 P102 P103 P104 P105 P106 P107 P108 P109 P110 P111 P112 P113 P114 P115 P116 P117 P118 P119 P120 P121 P122 P123 P124 P125 P126 P127 P128 P129 P130 P131 P132 P133 P134 P135 P136 P137 P138 P139 P140 P141 P142 P143 P144 P145 P146 P147 P148 P149 P150 P151 P152 P153 P154 P155 P156 P157 P158 P159 P160 P161 P162 P163 P164 P165 P166 P167 P168 P169 P170 P171 P172 P173 P174 P175 P176 P177 P178 P179 P180 P181 P182 P183 P184 P185 P186 P187 P188 P189 P190 P191 P192 P193 P194 P195 P196 P197 P198 P199 P200 P201 P202 P203 P204 P205 P206 P207 P208 P209 P210 P211 P212 P213 P214 P215 P216 P217 P218 P219 P220 P221 P222 P223 P224 P225 P226 P227 P228 P229 P230 P231 P232 P233 P234 P235 P236 P237 P238 P239 P240 P241 P242 P243 P244 P245 P246 P247 P248 P249 P250 P251 P252 P253 P254 P255 P256 P257 P258 P259 P260 P261 P262 P263 P264 P265 P266 P267 P268 P269 P270 P271 P272 P273 P274 P275 P276 P277 P278 P279 P280 P281 P282 P283 P284 P285 P286 P287 P288 P289 P290 P291 P292 P293 P294 P295 P296 P297 P298 P299 P300 P301 P302 P303 P304 P305 P306 P307 P308 P309 P310 P311 P312 P313 P314 P315 P316 P317 P318 P319 P320 P321 P322 P323 P324 P325 P326 P327 P328 P329 P330 P331 P332 P333 P334 P335 P336 P337 P338 P339 P340 P341 P342 P343 P344 P345 P346 P347 P348 P349 P350 P351 P352 P353 P354 P355 P356 P357 P358 P359 P360 P361 P362 P363 P364 P365 P366 P367 P368 P369 P370 P371 P372 P373 P374 P375 P376 P377 P378 P379 P380 P381 P382 P383 P384 P385 P386 P387 P388 P389 P390 P391 P392 P393 P394 P395 P396 P397 P398 P399 P400 P401 P402 P403 P404 P405 P406 P407 P408 P409 P410 P411 P412 P413 P414 P415 P416 P417 P418 P419 P420 P421 P422 P423 P424 P425 P426 P427 P428 P429 P430 P431 P432 P433 P434 P435 P436 P437 P438 P439 P440 P441 P442 P443 P444 P445 P446 P447 P448 P449 P450 P451 P452 P453 P454 P455 P456 P457 P458 P459 P460 P461 P462 P463 P464 P465 P466 P467 P468 P469 P470 P471 P472 P473 P474 P475 P476 P477 P478 P479 P480 P481 P482 P483 P484 P485 P486 P487 P488 P489 P490 P491 P492 P493 P494 P495 P496 P497 P498 P499 P500 P501 P502 P503 P504 P505 P506 P507 P508 P509 P510 P511 P512 P513 P514 P515 P516 P517 P518 P519 P520 P521 P522 P523 P524 P525 P526 P527 P528 P529 P530 P531 P532 P533 P534 P535 P536 P537 P538 P539 P540 P541 P542 P543 P544 P545 P546 P547 P548 P549 P550 P551 P552 P553 P554 P555 P556 P557 P558 P559 P560 P561 P562 P563 P564 P565 P566 P567 P568 P569 P570 P571 P572 P573 P574 P575 P576 P577 P578 P579 P580 P581 P582 P583 P584 P585 P586 P587 P588 P589 P590 P591 P592 P593 P594 P595 P596 P597 P598 P599 P600 P601 P602 P603 P604 P605 P606 P607 P608 P609 P610 P611 P612 P613 P614 P615 P616 P617 P618 P619 P620 P621 P622 P623 P624 P625 P626 P627 P628 P629 P630 P631 P632 P633 P634 P635 P636 P637 P638 P639 P640 P641 P642 P643 P644 P645 P646 P647 P648 P649 P650 P651 P652 P653 P654 P655 P656 P657 P658 P659 P660 P661 P662 P663 P664 P665 P666 P667 P668 P669 P670 P671 P672 P673 P674 P675 P676 P677 P678 P679 P680 P681 P682 P683 P684 P685 P686 P687 P688 P689 P690 P691 P692 P693 P694 P695 P696 P697 P698 P699 P700 P701 P702 P703 P704 P705 P706 P707 P708 P709 P710 P711 P712 P713 P714 P715 P716 P717 P718 P719 P720 P721 P722 P723 P724 P725 P726 P727 P728 P729 P730 P731 P732 P733 P734 P735 P736 P737 P738 P739 P740 P741 P742 P743 P744 P745 P746 P747 P748 P749 P750 P751 P752 P753 P754 P755 P756 P757 P758 P759 P760 P761 P762 P763 P764 P765 P766 P767 P768 P769 P770 P771 P772 P773 P774 P775 P776 P777 P778 P779 P780 P781 P782 P783 P784 P785 P786 P787 P788 P789 P790 P791 P792 P793 P794 P795 P796 P797 P798 P799 P800 P801 P802 P803 P804 P805 P806 P807 P808 P809 P810 P811 P812 P813 P814 P815 P816 P817 P818 P819 P820 P821 P822 P823 P824 P825 P826 P827 P828 P829 P830 P831 P832 P833 P834 P835 P836 P837 P838 P839 P840 P841 P842 P843 P844 P845 P846 P847 P848 P849 P850 P851 P852 P853 P854 P855 P856 P857 P858 P859 P860 P861 P862 P863 P864 P865 P866 P867 P868 P869 P870 P871 P872 P873 P874 P875 P876 P877 P878 P879 P880 P881 P882 P883 P884 P885 P886 P887 P888 P889 P890 P891 P892 P893 P894 P895 P896 P897 P898 P899 P900 P901 P902 P903 P904 P905 P906 P907 P908 P909 P910 P911 P912 P913 P914 P915 P916 P917 P918 P919 P920 P921 P922 P923 P924 P925 P926 P927 P928 P929 P930 P931 P932 P933 P934 P935 P936 P937 P938 P939 P940 P941 P942 P943 P944 P945 P946 P947 P948 P949 P950 P951 P952 P953 P954 P955 P956 P957 P958 P959 P960 P961 P962 P963 P964 P965 P966 P967 P968 P969 P970 P971 P972 P973 P974 P975 P976 P977 P978 P979 P980 P981 P982 P983 P984 P985 P986 P987 P988 P989 P990 P991 P992 P993 P994 P995 P996 P997 P998 P999 P1000 P1001 P1002 P1003 P1004 P1005 P1006 P1007 P1008 P1009 P1010 P1011 P1012 P1013 P1014 P1015 P1016 P1017 P1018 P1019 P1020 P1021 P1022 P1023 P1024 P1025 P1026 P1027 P1028 P1029 P1030 P1031 P1032 P1033 P1034 P1035 P1036 P1037 P1038 P1039 P1040 P1041 P1042 P1043 P1044 P1045 P1046 P1047 P1048 P1049 P1050 P1051 P1052 P1053 P1054 P1055 P1056 P1057 P1058 P1059 P1060 P1061 P1062 P1063 P1064 P1065 P1066 P1067 P1068 P1069 P1070 P1071 P1072 P1073 P1074 P1075 P1076 P1077 P1078 P1079 P1080 P1081 P1082 P1083 P1084 P1085 P1086 P1087 P1088 P1089 P1090 P1091 P1092 P1093 P1094 P1095 P1096 P1097 P1098 P1099 P1100 P1101 P1102 P1103 P1104 P1105 P1106 P1107 P1108 P1109 P1110 P1111 P1112 P1113 P1114 P1115 P1116 P1117 P1118 P1119 P1120 P1121 P1122 P1123 P1124 P1125 P1126 P1127 P1128 P1129 P1130 P1131 P1132 P1133 P1134 P1135 P1136 P1137 P1138 P1139 P1140 P1141 P1142 P1143 P1144 P1145 P1146 P1147 P1148 P1149 P1150 P1151 P1152 P1153 P1154 P1155 P1156 P1157 P1158 P1159 P1160 P1161 P1162 P1163 P1164 P1165 P1166 P1167 P1168 P1169 P1170 P1171 P1172 P1173 P1174 P1175 P1176 P1177 P1178 P1179 P1180 P1181 P1182 P1183 P1184 P1185 P1186 P1187 P1188 P1189 P1190 P1191 P1192 P1193 P1194 P1195 P1196 P1197 P1198 P1199 P1200 P1201 P1202 P1203 P1204 P1205 P1206 P1207 P1208 P1209 P1210 P1211 P1212 P1213 P1214 P1215 P1216 P1217 P1218 P1219 P1220 P1221 P1222 P1223 P1224 P1225 P1226 P1227 P1228 P1229 P1230 P1231 P1232 P1233 P1234 P1235 P1236 P1237 P1238 P1239 P1240 P1241 P1242 P1243 P1244 P1245 P1246 P1247 P1248 P1249 P1250 P1251 P1252 P1253 P1254 P1255 P1256 P1257 P1258 P1259 P1260 P1261 P1262 P1263 P1264 P1265 P1266 P1267 P1268 P1269 P1270 P1271 P1272 P1273 P1274 P1275 P1276 P1277 P1278 P1279 P1280 P1281 P1282 P1283 P1284 P1285 P1286 P1287 P1288 P1289 P1290 P1291 P1292 P1293 P1294 P1295 P1296 P1297 P1298 P1299 P1300 P1301 P1302 P1303 P1304 P1305 P1306 P1307 P1308 P1309 P1310 P1311 P1312 P1313 P1314 P1315 P1316 P1317 P1318 P1319 P1320 P1321 P1322 P1323 P1324 P1325 P1326 P1327 P1328 P1329 P1330 P1331 P1332 P1333 P1334 P1335 P1336 P1337 P1338 P1339 P1340 P1341 P1342 P1343 P1344 P1345 P1346 P1347 P1348 P1349 P1350 P1351 P1352 P1353 P1354 P1355 P1356 P1357 P1358 P1359 P1360 P1361 P1362 P1363 P1364 P1365 P1366 P1367 P1368 P1369 P1370 P1371 P1372 P1373 P1374 P1375 P1376 P1377 P1378 P1379 P1380 P1381 P1382 P1383 P1384 P1385 P1386 P1387 P1388 P1389 P1390 P1391 P1392 P1393 P1394 P1395 P1396 P1397 P1398 P1399 P1400 P1401 P1402 P1403 P1404 P1405 P1406 P1407 P1408 P1409 P1410 P1411 P1412 P1413 P1414 P1415 P1416 P1417 P1418 P1419 P1420 P1421 P1422 P1423 P1424 P1425 P1426 P1427 P1428 P1429 P1430 P1431 P1432 P1433 P1434 P1435 P1436 P1437 P1438 P1439 P1440 P1441 P1442 P1443 P1444 P1445 P1446 P1447 P1448 P1449 P1450 P1451 P1452 P1453 P1454 P1455 P1456 P1457 P1458 P1459 P1460 P1461 P1462 P1463 P1464 P1465 P1466 P1467 P1468 P1469 P1470 P1471 P1472 P1473 P1474 P1475 P1476 P1477 P1478 P1479 P1480 P1481 P1482 P1483 P1484 P1485 P1486 P1487 P1488 P1489 P1490 P1491 P1492 P1493 P1494 P1495 P1496 P1497 P1498 P1499 P1500 P1501 P1502 P1503 P1504 P1505 P1506 P1507 P1508 P1509 P1510 P1511 P1512 P1513 P1514 P1515 P1516 P1517 P1518 P1519 P1520 P1521 P1522 P1523 P1524 P1525 P1526 P1527 P1528 P1529 P1530 P1531 P1532 P1533 P1534 P1535 P1536 P1537 P1538 P1539 P1540 P1541 P1542 P1543 P1544 P1545 P1546 P1547 P1548 P1549 P1550 P1551 P1552 P1553 P1554 P1555 P1556 P1557 P1558 P1559 P1560 P1561 P1562 P1563 P1564 P1565 P1566 P1567 P1568 P1569 P1570 P1571 P1572 P1573 P1574 P1575 P1576 P1577 P1578 P1579 P1580 P1581 P1582 P1583 P1584 P1585 P1586 P1587 P1588 P1589 P1590 P1591 P1592 P1593 P1594 P1595 P1596 P1597 P1598 P1599 P1600 P1601 P1602 P1603 P1604 P1605 P1606 P1607 P1608 P1609 P1610 P1611 P1612 P1613 P1614 P1615 P1616 P1617 P1618 P1619 P1620 P1621 P1622 P1623 P1624 P1625 P1626 P1627 P1628 P1629 P1630 P1631 P1632 P1633 P1634 P1635 P1636 P1637 P1638 P1639 P1640 P1641 P1642 P1643 P1644 P1645 P1646 P1647 P1648 P1649 P1650 P1651 P1652 P1653 P1654 P1655 P1656 P1657 P1658 P1659 P1660 P1661 P1662 P1663 P1664 P1665 P1666 P1667 P1668 P1669 P1670 P1671 P1672 P1673 P1674 P1675 P1676 P1677 P1678 P1679 P1680 P1681 P1682 P1683 P1684 P1685 P1686 P1687 P1688 P1689 P1690 P1691 P1692 P1693 P1694 P1695 P1696 P1697 P1698 P1699 P1700 P1701 P1702 P1703 P1704 P1705 P1706 P1707 P1708 P1709 P1710 P1711 P1712 P1713 P1714 P1715 P1716 P1717 P1718 P1719 P1720 P1721 P1722 P1723 P1724 P1725 P1726 P1727 P1728 P1729 P1730 P1731 P1732 P1733 P1734 P1735 P1736 P1737 P1738 P1739 P1740 P1741 P1742 P1743 P1744 P1745 P1746 P1747 P1748 P1749 P1750 P1751 P1752 P1753 P1754 P1755 P1756 P1757 P1758 P1759 P1760 P1761 P1762 P1763 P1764 P1765 P1766 P1767 P1768 P1769 P1770 P1771 P1772 P1773 P1774 P1775 P1776 P1777 P1778 P1779 P1780 P1781 P1782 P1783 P1784 P1785 P1786 P1787 P1788 P1789 P1790 P1791 P1792 P1793 P1794 P1795 P1796 P1797 P1798 P1799 P1800 P1801 P1802 P1803 P1804 P1805 P1806 P1807 P1808 P1809 P1810 P1811 P1812 P1813 P1814 P1815 P1816 P1817 P1818 P1819 P1820 P1821 P1822 P1823 P1824 P1825 P1826 P1827 P1828 P1829 P1830 P1831 P1832 P1833 P1834 P1835 P1836 P1837 P1838 P1839 P1840 P1841 P1842 P1843 P1844 P1845 P1846 P1847 P1848 P1849 P1850 P1851 P1852 P1853 P1854 P1855 P1856 P1857 P1858 P1859 P1860 P1861 P1862 P1863 P1864 P1865 P1866 P1867 P1868 P1869 P1870 P1871 P1872 P1873 P1874 P1875 P1876 P1877 P1878 P1879 P1880 P1881 P1882 P1883 P1884 P1885 P1886 P1887 P1888 P1889 P1890 P1891 P1892 P1893 P1894 P1895 P1896 P1897 P1898 P1899 P1900 P1901 P1902 P1903 P1904 P1905 P1906 P1907 P1908 P1909 P1910 P1911 P1912 P1913 P1914 P1915 P1916 P1917 P1918 P1919 P1920 P1921 P1922 P1923 P1924 P1925 P1926 P1927 P1928 P1929 P1930 P1931 P1932 P1933 P1934 P1935 P1936 P1937 P1938 P1939 P1940 P1941 P1942 P1943 P1944 P1945 P1946 P1947 P1948 P1949 P1950 P1951 P1952 P1953 P1954 P1955 P1956 P1957 P1958 P1959 P1960 P1961 P1962 P1963 P1964 P1965 P1966 P1967 P1968 P1969 P1970 P1971 P1972 P1973 P1974 P1975 P1976 P1977 P1978 P1979 P1980 P1981 P1982 P1983 P1984 P1985 P1986 P1987 P1988 P1989 P1990 P1991 P1992 P1993 P1994 P1995 P1996 P1997 P1998 P1999 P2000 P2001 P2002 P2003 P2004 P2005 P2006 P2007 P2008 P2009 P2010 P2011 P2012 P2013 P2014 P2015 P2016 P2017 P2018 P2019 P2020 P2021 P2022 P2023 P2024 P2025 P2026 P2027 P2028 P2029 P2030 P2031 P2032 P2033 P2034 P2035 P2036 P2037 P2038 P2039 P2040 P2041 P2042 P2043 P2044 P2045 P2046 P2047 P2048 P2049 P2050 P2051 P2052 P2053 P2054 P2055 P2056 P2057 P2058 P2059 P2060 P2061 P2062 P2063 P2064 P2065 P2066 P2067 P2068 P2069 P2070 P2071 P2072 P2073 P2074 P2075 P2076 P2077 P2078 P2079 P2080 P2081 P2082 P2083 P2084 P2085 P2086 P2087 P2088 P2089 P2090 P2091 P2092 P2093 P2094 P2095 P2096 P2097 P2098 P2099 P2100 P2101 P2102 P2103 P2104 P2105 P2106 P2107 P2108 P2109 P2110 P2111 P2112 P2113 P2114 P2115 P2116 P2117 P2118 P2119 P2120 P2121 P2122 P2123 P2124 P2125 P2126 P2127 P2128 P2129 P2130 P2131 P2132 P2133 P2134 P2135 P2136 P2137 P2138 P2139 P2140 P2141 P2142 P2143 P2144 P2145 P2146 P2147 P2148 P2149 P2150 P2151 P2152 P2153 P2154 P2155 P2156 P2157 P2158 P2159 P2160 P2161 P2162 P2163 P2164 P2165 P2166 P2167 P2168 P2169 P2170 P2171 P2172 P2173 P21

	Column Name	Data Type	Nullable	Default	Description
					Part of the datab service key.
	fabric_service_uri	nvarchar(256)	True		The URI of the database service Combined with fabric cluster provides the fore key to the datab service.

Constraints:

Unique (SQLInstanceName & Name)

Logical Database Features

Table: tbl_sql_database_features

View: sql_database_features

FSM: DatabaseFeatureStateMachine

Contains one row per feature enabled on a database. Used for allow_listing features on a specific database.

Name	Data Type	Nullable	Default	Description
database_id	bigint	False		The id of the database having the feature.
feature_name	nvarchar(50)	False		The name of the feature being described.
feature_value	sql_variant	False		The feature value for this database.

GeoDR Relationships [was GeoDR Relations]

Table: tbl_geodr_relationships

View: geodr_relationships

FSM: GeodrRelationshipStateMachine

Contains one row per GeoDR relationship. Each row describes a replication relationship between a source database defined in the SQL Cluster and another database that may be in the current SQL cluster or some other SQL Cluster.

Note: Placeholder table schema is based on a modified version of the SAWA v1 Gateway Metadata schema and is subject to review.

	Column Name	Data Type	Nullable	Default	Description
PK	source_database_id	bigint	False		
PK	replica_sql_cluster_name	nvarchar(128)	False		The SQL cluster in which the replica is located. May be a reference to the current or different SQL cluster.
PK	replica_database_guid	bigint	False		
	link_id	uniqueidentifier	False		
	geo_replica_state	nvarchar(64)	True		
	geo_replica_substate	nvarchar(64)	True		
	copy_complete_subscription_id	uniqueidentifier	False		
	terminate_complete_subscription_id	uniqueidentifier	False		
	is_created_with_database	bit	False	0	

Logical Servers

Table: tbl_logical_servers

View: logical_servers

FSM: LogicalServerStateMachine

Contains one row per customer (logical) server.

	Column Name	Data Type	Nullable	Default	Description
PK	name	nvarchar(128)	False		Logical server name.
	logical_server_id	uniqueidentifier	False	newsequentialid()	A guid assigned to the
	sql_instance_group [was InstanceGroupName]	nvarchar(128)	True		The SQL instance group that contains the SQL instances that host the databases in this logical server.
	control_ring_dns_name	nvarchar(256)	True		<verify usage, nullable?>
	customer_subscription_id [new]	uniqueidentifier	False		The SQL customer subscription Id under which this logical server was created.
	admin_login_name [was AdminName]	nvarchar(128)	True		The admin user login name assigned by the customer. A temporarily stored value, only

	Column Name	Data Type	Nullable	Default	Description
					stored while being applied to the database.
	admin_login_password [was AdminPassword]	varbinary(1024)	True		The admin password. Encrypted. A temporarily stored value, only stored while being applied to the database.

SQL Cluster Infrastructure

SQL Clusters [was Global Metadata]

Table: tbl_sql_clusters

View: sql_clusters

FSM: SqlClusterStateMachine

Contains one row representing the SQL cluster itself. Contains metadata that is either specific to the SQL cluster or is global to the SQL Azure environment.

	Column Name	Data Type	Nullable	Description
PK	name	nvarchar(128)	False	The SQL cluster name.
	primary_control_ring_name [was PrimaryControlClusterName]	nvarchar(128)	False	
	add_remove_control_ring_sequence	bigint	False	
	security_principal_sequence	Bigint	False	

SQL Control Rings [was Control Hosted Services]

Table: tbl_sql_control_rings extends tbl_fabric_clusters

View: sql_control_rings

FSM: SqlControlRingStateMachine

Contains one row per control ring in the SQL cluster.

Initially only one control ring will exist per SQL Cluster thus only one per row per CMS database. The North Star architecture allows for multiple control rings to exist in a single SQL cluster. A primary ring contains the active CMS which is the only stateful service on the control ring. Other control ring services can be accessed at either ring. This provides protection in the case that a stateless service in the ring is unavailable or overloaded.

	Column Name	Data Type	Nullable	Description
PK	name [was ClusterName]	nvarchar(128)	False	Control ring name. This is the DNS name of the fabric cluster that forms the control ring. <example>
	add_remove_sequence	bigint	True	
	processed_alias_uri	nvarchar(256)	True	
	dns_record_id	uniqueidentifier	True	
	is_bootstrap	bit	False	
	ipv4_address	nvarchar(15)	True	

SQL Instances [was Instances]

Table: tbl_sql_instances

View: sql_instances

FSM: SqlInstanceStateMachine

Contains one row per SQL Server instance running in a tenant ring and composed within a SQL Instance Group. Includes both Remote Storage and Local Storage instances.

	Column Name	Data Type	Nullable	Description
PK	name	nvarchar(128)	False	<p>The name of the SQL Instance, constructed from the logical server name + a suffix indicating either the hosted premium database or a count of SQL instances in the SQL Instance group associated with the logical server, where 0 indicates this is the master instance containing the logical master database for a logical server.</p> <p>Examples H1679ab12\0 – the master SQL instance for a logical server H1679ab12<database guid> - an instance containing a premium database</p>
	sql_instance_id [was Uniqueid]	uniqueidentifier	False	
	Type	nvarchar(128)	False	<p>The type of SQL instance</p> <p>CMS SQL.LocalStorage SQL.RemoteStorage FQS.Compute FQS.Control</p>
	service_level_objective	nvarchar(128)	False	<p>The service level objective assigned to the instance which acts as a size for the instance and determines the way the Windows Fabric</p>

	Column Name	Data Type	Nullable	Description
				allocates instances to a node. Values are: Shared P1 P2 P2 P4
	target_service_level_objective	nvarchar(128)	True	The target service level objective, used when changing the size of the instance. Values are:Shared P1 P2 P2 P4
	worker_service_affinity_tag	nvarchar(128)	True	<Is this a general purpose affinity tag that should be in fabric_applications?>
	storage_account_affinity_tag	nvarchar(128)	True	
	sql_instance_group_name [was InstanceGroupName]	nvarchar(128)	True	The SQL Instance Group in which this SQL instance participates.
	security_principal_id	bigint	True	
	collation	nvarchar(128)	True	
	trace_flags	xml	True	
	reset_admin_login_name [was ResetAdminName]	nvarchar(128)	True	
	reset_admin_login_password [was ResetAdminPasswordHash]	varbinary(1024)	True	
	reset_admin_login_time [was ResetAdminDate]	datetime2	True	
	tenant_ring_name [was ClusterName]	nvarchar(128)	True	Local fabric cluster DNS name (bound to a Worker service) <clarify> <verify>

	Column Name	Data Type	Nullable	Description
				column length vs. DNS name>
	fabric_application_uri	nvarchar(256)	True	With tenant_ring_name identifies the application that implements this SQL Instance.
	dns_name	nvarchar(256)	True	Global DNS (bound to Control services) <clarify>
	fabric_cluster_dns_name [was ClusterDnsName]	nvarchar(256)	True	<clarify wrt FabricClusterName, DNS name>
	fabric_service_uri	nvarchar(256)	True	Current Winfab service Uri
	fabric_service_sequence	bigint	True	
	storage_account_dns_name	nvarchar(128)	True	Current placement
	storage_container_name	nvarchar(64)	True	
	storage_policy_name	nvarchar(128)	True	
	master_key_alg	nvarchar(32)	True	
	master_key_length	int	True	
	master_key	[was MasterKeyBlob]	varbinary(MAX)	True
	processed_database_name	nvarchar(128)	True	Transient reference used during drop or other ordered workflow processing over multiple databases.

	Column Name	Data Type	Nullable	Description
	extension_data	xml	True	DW sourced <not here?>

SQL Instance Groups [was Instance Groups]

Table: tbl_sql_instance_groups

View: sql_instance_groups

FSM: SqlInstanceStateMachine

Contains one row per SQL Instance Group. Includes a row for CMS in the control ring to support holding storage group and keys.

Column Name	Data Type	Nullable	Description
name	nvarchar(128)	False	The name of the SQL Instance Group. For SQL database is either 'CMS' or the corresponding logical server name.
members [delete]	xml	True	Used to manage locking the set of databases in the absence of an efficient query and lock behavior for a set of state machines in FSM
security_principal_id	bigint	True	
security_principal_password [was Password]	varbinary(1024)	True	
extension_data	xml	True	Reserved for FQS use only. May be revised later.

SQL Tenant Ring [was Worker Hosted Services]

Table: tbl_sql_tenant_rings

View: sql_tenant_rings

FSM: SqlTenantRingStateMachine

Contains one row per tenant ring in the SQL cluster. Extends tbl_fabric_hosted_services.

There may be multiple tenant rings per SQL cluster.

Column Name	Data Type	Nullable	Description
name [was ClusterName]	nvarchar(128)	False	The tenant ring name. This is the name of the fabric cluster that forms the control ring.<example>
certificate	varbinary(MAX)	True	Certificate used for ring-wide encryption of data. Not used in OneBox implementations.
initial_counts	xml	True	Initial allocation of nodes per node role in the tenant ring. Will be used to assist in allocation of nodes to VMs during placement. Initially, SQL DDC will use the initial values as the max.
current_counts	xml	True	Current count of nodes of each role.
used_counts	xml	True	
target_counts	xml	True	Target count. Used to allow buffering for expansion. The tenant ring state machine will grow the ring or shrink the ring as required if the target counts differ from the current counts.
placement_base_weight	int	False	
placement_affinity_tag	nvarchar(128)	True	A tag used in placement expressions to ensure only control ring applications are placed on the control ring.

Constraints:

Secrets

Table: tbl_secrets

View: secrets

FSM: SecretsStateMachine

Contains one row per individual secret.

Column Name	Data Type	Nullable	Description
consumer_name	nvarchar(128)	False	Name of the consumer, could be instance or instance group
consumer_type	nvarchar(128)	False	Consumer type
secret_type	nvarchar(128)	False	Type of Secret. Example: Cert, SAS key, Symmetric key, etc.
generation_id	int	False	Generation or version of the secret
state	nvarchar(128)	False	Indicates state of the FSM
thumbprint	nvarchar(128)	True	Primary generation number. Used during rollover of secret
public_blob	nvarchar(MAX)	True	Public key of the blob (could be null for symmetric keys)
private_blob	Nvarchar(MAX)	False	Private key blob
issuer_name	nvarchar(128)	False	Who issued this key – to track changes to issuer
issuer_type	int	False	Type of issuer
properties	XML	True	Customer properties for this secret (e.g. key size, expiration, policy or container name for SAS keys, etc.)
marked_as_deleted	Bit	False	A bit is set if the secret should be deleted from clients

Instance Group Secrets

Table: tbl_instance_group_secrets

View: instance_group_secrets

FSM: InstanceGroupSecretsStateMachine

Contains one row per instance group secret.

Column Name	Data Type	Nullable	Description
name	nvarchar(128)	False	Name of the instance group
state	nvarchar(128)	False	Indicates state of the FSM
previous_generation	int	True	Previous generation number. Used during rollover of secret
current_generation	int	False	Current generation number.
next_generation	int	True	Next generation number. Used during rollover of secret
primary_generation	int	False	Primary generation number. Used during rollover of secret

Instance Secrets

Table: tbl_instance secrets

View: instance secrets

FSM: InstanceSecretsStateMachine

Contains one row per instance secret.

Column Name	Data Type	Nullable	Description
name	nvarchar(16)	False	Name of the instance
state	nvarchar (128)	False	Indicates state of the FSM
previous_generation	int	True	Previous generation number. Used during rollover of secret
current_generation	int	False	Current generation number.
next_generation	int	True	Next generation number. Used during rollover of secret
primary_generation	int	False	Primary generation number. Used during rollover of secret
current_app_property	nvarchar(128)	True	Current App property name

Infrastructure

Azure Storage Accounts

Table: tbl_azure_storage_accounts

View: azure_storage_accounts

FSM: StorageAccountStateMachine

Contains one row per storage account.

	Column Name	Data Type	Nullable	Description
PK	dns_name	nvarchar(128)	False	account name, (e.g. tomtalbay.blob.core.windows.net)
	subscription_id	nvarchar(128)	False	
	storage_account_name	nvarchar(128)	False	Account name of subscription
	containers_used	int	False	Number of containers used.
	primary_key	varbinary(4000)	True	Primary account key
	secondary_key	varbinary(4000)	True	Secondary account key
	use_primary_key	bit	False	Indicates if the primary key should be used for SAS signatures.
	placement_base_weight	int	False	Weight used for allocation.
	placement_affinity_tag	nvarchar(128)	True	Affinity tag.

Constraints:

Unique: (subscription_id & storage_account_name)

Azure DNS Records

Table: tbl_fabric_aliases

View: fabric_aliases

FSM: FabricAliasStateMachine

Contains one row per DNS Record.

TBD

Fabric Aliases

Table: tbl_fabric_aliases

View: fabric_aliases

FSM: FabricAliasStateMachine

Contains one row per Windows Fabric alias.

	Column Name	Data Type	Nullable	Description
PK	fabric_service_uri	nvarchar(256)	False	The exposed alias URI.
	add_remove_control_service_sequence	bigint	False	Sequence number for add remove events
	target_fabric_cluster_name	nvarchar(128)	False	The cluster name of the target fabric service.
	target_fabric_service_uri	nvarchar(256)	False	The service URI component of the underlying fabric service identifier.

Fabric Application

Table: tbl_fabric_applications

View: fabric_applications

FSM: FabricApplicationMachine

Contains a row for each Fabric Application.

	Column Name	Data Type	Nullable	Description
PK	fabric_cluster_name [was ClusterName]	nvarchar(128)	False	The name of the fabric cluster which this application is located on.
PK	fabric_application_uri	nvarchar(256)	False	
	deployment_id_sequence	bigint	True	Sequence number used to track progress of a redeployment or interruption.
	application_type_name	nvarchar(256)	False	Name of the fabric application.
	current_application_type_version	nvarchar(256)	True	The current fabric application version of this application. It is null if the application has not been created.
	target_application_type_version	nvarchar(256)	False	The target fabric application version. This is different from current version if there is an upgrade in progress, or if there is an upgrade in progress that has failed.
	deployment_parameters	xml	True	Parameter values that are substituted into the application manifest during windows fabric application provisioning. Parameters used vary by application type.
	upgrade_policy	xml	True	Transient state during upgrade. Upgrade policy to use when WinFab to kick off the upgrade. Essentially XML-serialized UpgradePolicyDescription.
	rollout_key	nvarchar(256)	True	Transient state during upgrade of the currently in-progress rollout for this app instance.
	last_upgrade_start_time	datetime2	True	Start time of the last upgrade of the application instance.

	Column Name	Data Type	Nullable	Description
	last_upgrade_progress_change_time	Datetime2	True	Last time the upgrade prog changed for this applicator instance.
	upgrade_progress	xml	True	Xml serialization of WinFabApplicationUpgrade for this application instance



Fabric Application Settings

Table: tbl_fabric_application_settings [TBD]

View: fabric_application_settings [TBD]

FSM: FabricApplicationTypeMachine

Contains deployment configuration settings for a specific application type.

Schema TBD, likely to be structured as multiple tables that are collectively related to tbl_application_type and **not** managed by a state machine.

Fabric Application Type

Table: tbl_fabric_application_types

View: fabric_application_types

FSM: FabricApplicationTypeMachine

Contains a row for each Fabric Application Type for each fabric cluster and additional state to drive/track the progress of an application type upgrade.

	Column Name	Data Type	Nullable	Description
PK	fabric_cluster_name	nvarchar(128)	False	The name of the fabric cluster on which this application type is located.
PK	application_type_name	nvarchar(256)	False	Name of fabric application type.
	application_build_path	nvarchar(256)	True	Relative path in the image store of the package we're currently upgrading to.
	current_application_type_version	nvarchar(256)	False	The current version of this application type on this fabric cluster. This is the version new applications will get instantiated with.
	target_application_type_version	nvarchar(256)	False	Target version for this application type on this cluster. This is different from the current version only when there is an upgrade going on.
	applications_upgrading	nvarchar(max)	True	Pipe () separated list of applications currently undergoing upgrade that are being tracked. [will be upgrade to an xml column in a later change]
	last_processed_application	nvarchar(256)	True	Uri of the last application upgrade was kicked off for. Used to save iteration state when kicking off upgrade for a large number of applications.
	deployment_tag	nvarchar(256)	True	Textual tag for the current deployment going on.
	rollout_key	nvarchar(256)	True	Transient state during upgrade. Key of the

	Column Name	Data Type	Nullable	Description
				currently in-progress rollout for this app type.
	upgrade_policy	xml	True	Transient state during upgrade. Upgrade policy to use when calling WinFab to kick off the upgrade. Essentially XML-serialized UpgradePolicyDescription.
	last_rollout_outcome	bit	True	Indicates the overall success or failure for the whole rollout. Starts out as true, flips to false as soon as an app instance fails upgrade.
	last_rollout_start_time	datetime2	True	Start time of the last rollout for this application type.

Fabric Cluster [was Fabric Hosted Service]

Table: tbl_fabric_clusters

View: fabric_clusters

FSM: FabricClusterStateMachine

Contains a row for each Fabric Cluster within the SQL Cluster. Fabric Cluster contains base data common to SQL Control Rings and SQL Tenant Rings.

	Column Name	Data Type	Nullable	Description
PK	name [was ClusterName]	nvarchar(128)	False	
	hosted_service_subscription_id	nvarchar(128)	False	The internal Microsoft-owned subscription id under which the fabric cluster is created.
	hosted_service_name	nvarchar(128)	False	The exposed Azure name of the fabric cluster.
	hosted_service_dns_name	nvarchar(128)	True	<verify usage, name; is this different to the above>
	deployment_id	nvarchar(128)	True	
	deployment_id_sequence	bigint	True	Sequence number used to track the progress of a deployment in case of interruption. <verify usage>
	deployment_id_replaced_sequence	bigint	True	Sequence number used to track the progress of a redeployment in case of interruption.
	processed_fabric_uri	nvarchar(256)	True	<verify usage>
	ipv4_address	nvarchar(128)	True	

Fabric Services

Table: tbl_fabric_services

View: fabric_services

FSM: FabricServiceStateMachine

Contains a row for each Fabric Service deployed in the SQL Cluster. Includes database services and SQL instance application services deployed on the tenant rings, as well as infrastructural services deployed to both the control ring and tenant rings.

Column Name	Data Type	Nullable	Description
fabric_cluster_name [was ClusterName]	nvarchar(128)	False	The fabric cluster name.
fabric_service_uri	nvarchar(256)	False	
deployment_id_sequence	bigint	True	
application_uri	nvarchar(256)	False	
service_type_name	nvarchar(256)	False	
placement_constraints	nvarchar(2000)	True	
initialization_data	varbinary(MAX)	True	
kind	nvarchar(128)	False	Stateless or Stateful
target_replica_set_size	int	False	
has_persisted_state	bit	True	An additional qualifier required by Windows Fabric when kind == Stateful
min_replica_set_size	int	True	
metrics	xml	True	
correlations	xml	True	

Fabric Nodes

Table: tbl_fabric_nodes

View: fabric_nodes

FSM: NodeStateMachine

Contains a row for each Fabric Node in the Fabric Clusters. Key to uniquely identify a node is the combination of node name and the fabric cluster name it belongs to.

Column Name	Data Type	Nullable	Description
name	nvarchar(128)	False	The node name.
fabric_cluster_name	nvarchar(128)	False	Name of the fabric cluster the node belongs to.
fabric_state	nvarchar(128)	False	Textual representation of the state reported by WinFabric. This gets refreshed periodically (currently every 1 minute).
deactivation_intent	nvarchar(64)	True	The intent expressed for deactivation, if the state machine is in the process of deactivating the node. NULL otherwise.

How good have you found this content?

