Voice Microservices Private Edition Guide

Agent State Service metrics and alerts

7/14/2025

Contents

Find the metrics Agent State Service exposes and the alerts defined for Agent State Service.

Service	CRD or annotations?	Port	Endpoint/Selector	Metrics update interval
Agent State Service	PodMonitor	11000	http://:11000/metrics	30 seconds

See details about:

- Agent State Service metrics
- Agent State Service alerts

Metrics

Voice Agent State Service exposes Genesys-defined, Agent State Service–specific metrics as well as some standard Kafka metrics. You can query Prometheus directly to see all the metrics that the Agent State Service exposes. The following metrics are likely to be particularly useful. Genesys does not commit to maintain other currently available Agent State Service metrics not documented on this page.

Metric and description	Metric details	Indicator of
agent_redis_state Current Redis connection state: -1 - error 0 - disconnected 1 - connected 2 - ready	**Unit:** N/A **Type:** gauge **Label:** location, redis_cluster_name **Sample value:** 2	
agent_stream_redis_state Current Tenant Redis connection state: 0 - disconnected 1 - connected	**Unit:** N/A **Type:** gauge **Label:** location, redis_cluster_name **Sample value:** 1	
agent_total_sessions Total number of agent sessions.	**Unit:** N/A **Type:** gauge **Label:** tenant **Sample value:**	Saturation
agent_callevents Total number of received call events.	**Unit:** N/A **Type:** counter **Label:** tenant	Traffic

Metric and description	Metric details	Indicator of
	Sample value:	
agent_logged_in_agents Number of logged-in agents.	**Unit:** N/A **Type:** gauge **Label:** tenant **Sample value:**	Saturation
agent_health_level Health level of the agent node: -1 - error 0 - fail 1 - degraded 2 - pass	**Unit:** N/A **Type:** gauge **Label:** tenant **Sample value:** 2	Traffic
agent_envoy_proxy_status Status of the Envoy proxy: -1 - error 0 - disconnected 1 - connected	**Unit:** N/A **Type:** gauge **Label:** **Sample value:** 1	
agent_config_node_status Status of the config node connection: 0 - disconnected 1 - connected	**Unit:** N/A **Type:** gauge **Label:** **Sample value:** 1	
http_client_request_duration_seconds HTTP client time from request to response, in seconds.	**Unit:** seconds **Type:** histogram **Label:** target_service_name **Sample value:**	
http_client_response_count HTTP client responses received.	**Unit:** N/A **Type:** counter **Label:** target_service_name, tenant, status **Sample value:**	Traffic
kafka_consumer_recv_messages_total Number of messages received from Kafka.	**Unit:** N/A **Type:** counter **Label:** topic, tenant, kafka_location **Sample value:**	Traffic
kafka_consumer_error_total Number of Kafka consumer errors.	**Unit:** N/A **Type:** counter **Label:** topic, kafka_location **Sample value:**	Errors
kafka_consumer_latency Consumer latency is the time difference between when the message is produced	**Unit:** **Type:** histogram **Label:** topic, tenant, kafka_location	Latency

Metric and description	Metric details	Indicator of
and when the message is consumed. That is, the time when the consumer received the message minus the time when the producer produced the message.	**Sample value:**	
kafka_consumer_rebalance_total Number of Kafka consumer re-balance events.	**Unit:** N/A **Type:** counter **Label:** topic, kafka_location **Sample value:**	
kafka_consumer_state Current state of the Kafka consumer.	**Unit:** N/A **Type:** gauge **Label:** topic, kafka_location **Sample value:**	
kafka_producer__messages_total Number of messages received from Kafka.	**Unit:** N/A **Type:** counter **Label:** topic, tenant, kafka_location **Sample value:**	
kafka_producer_queue_depth Number of Kafka producer pending events.	**Unit:** N/A **Type:** gauge **Label:** kafka_location **Sample value:**	Saturation
kafka_producer_queue_age_seconds Age of the oldest producer pending event in seconds.	**Unit:** seconds **Type:** gauge **Label:** kafka_location **Sample value:**	
kafka_producer_error_total Number of Kafka producer errors.	**Unit:** N/A **Type:** counter **Label:** kafka_location **Sample value:**	
kafka_producer_state Current state of the Kafka producer.	**Unit:** N/A **Type:** gauge **Label:** kafka_location **Sample value:**	
log_output_bytes_total Total amount of log output, in bytes.	**Unit:** bytes **Type:** counter **Label:** level, format, module **Sample value:**	

Alerts

The following alerts are defined for Agent State Service.

Alert	Severity	Description	Based on	Threshold
Kafka events latency is too high	Warning	Actions: • If the alarm is triggered for multiple topics, ensure there are no issues with Kafka (CPU, memory, or network overload). • If the alarm is triggered only for topic {{ $labels.topic }}, check if there is an issue with the service related to the topic (CPU, memory, or network overload).	kafka_consumer_latency_bucket	Latency for more than 5% of messages is more than 0.5 seconds for topic {{ $labels.topic }}.
Possible messages lost	Critical	Actions: • Check Kafka and {{ $labels.job }} service overload, network degradation.	kafka_consumer_recv_messages_total, kafka_producer_sent_messages_total	Number of sent requests is two times higher than received for topic {{ $labels.topic }}.
Too many Kafka consumer failed health checks	Warning	Actions: • If the alarm is triggered for multiple services, ensure there are no issues with Kafka. Restart Kafka. • If the alarm is triggered only for container {{ $labels.container }}, check if there is an issue with the service.	kafka_consumer_error_total	Health check failed more than 10 times in 5 minutes for Kafka consumer for topic {{ $labels.topic}}.

Alert	Severity	Description	Based on	Threshold
Too many Kafka consumer request timeouts	Warning	Actions: • If the alarm is triggered for multiple services, ensure there are no issues with Kafka. Restart Kafka. • If the alarm is triggered only for container {{ $labels.container }}, check if there is an issue with the service.	kafka_consumer_error_total	More than 10 request timeouts appeared in 5 minutes for Kafka consumer for topic {{ $labels.topic}}.
Too many Kafka consumer crashes	Critical	Actions: • If the alarm is triggered for multiple services, ensure there are no issues with Kafka. Restart Kafka. • If the alarm is triggered only for container {{ $labels.container }}, check if there is an issue with the service.	kafka_consumer_error_total	More than 3 Kafka consumer crashes in 5 minutes for service {{ $labels.container }}.
Pod status Failed	Warning	Actions: • Restart the pod. Check if there are any issues with the pod after restart.	kube_pod_status_phase	Pod {{ $labels.pod }} is in Failed state.
Pod status Unknown	Warning	Actions: • Restart the pod. Check if	kube_pod_status_phase	Pod {{ $labels.pod }} is in Unknown state for 5 minutes.

Alert	Severity	Description	Based on	Threshold
		there are any issues with pod after restart.		
Pod status Pending	Warning	Actions: • Restart the pod. Check if there are any issues with the pod after restart.	kube_pod_status_phase	Pod {{ $labels.pod }} is in Pending state for 5 minutes.
Pod status NotReady	Critical	Actions: • Restart the pod. Check if there are any issues with the pod after restart.	kube_pod_status_ready	Pod {{ $labels.pod }} is in NotReady status for 5 minutes.
Container restarted repeatedly	Critical	Actions: • Check if the new version of the image was deployed. • Check for issues with the Kubernetes cluster.	kube_pod_container_status_restarts_total	Container {{ $labels.container }} was restarted 5 or more times within 15 minutes.
Max replicas is not sufficient for 5 mins	Critical	The desired number of replicas is higher than the current available replicas for the past 5 minutes.	kube_statefulset_replicas, kube_statefulset_status_replicas	The desired number of replicas is higher than the current available replicas for the past 5 minutes.
Kafka not available	Critical	Actions: • If the alarm is triggered for multiple services, ensure there are no issues with Kafka. Restart Kafka.	kafka_producer_state, kafka_consumer_state	Kafka is not available for pod {{ $labels.pod }} for 5 consecutive minutes.

Alert	Severity	Description	Based on	Threshold
		• If the alarm is triggered only for pod {{ $labels.pod }}, check if there is an issue with the pod.		
Redis not available	Critical	Actions: • If the alarm is triggered for multiple services, ensure there are no issues with Redis. Restart Redis. • If the alarm is triggered only for pod {{ $labels.pod }}, check if there is an issue with the pod.	agent_redis_state, agent_stream_redis_state	Redis is not available for pod {{ $labels.pod }} for 5 consecutive minutes.
Agent service fail	Critical	Actions: • Check if there is any problem with pod {{ $labels.pod }}, then restart the pod.	agent_health_level	Agent health level is Fail for pod {{ $labels.pod }} for 5 consecutive minutes.
Config node fail	Warning	Actions: • Check if there is any problem with pod {{ $labels.pod }} and the config node.	http_client_response_count	Requests to the config node fail for 5 consecutive minutes.
Pod CPU greater than 65%	Warning	High CPU load for pod {{ $labels.pod }}.	container_cpu_usage_seconds_total, container_spec_cpu_period	Container {{ $labels.container }} CPU usage exceeded 65% for 5 minutes.
Pod CPU greater	Critical	Critical CPU load	container_cpu_usage_seconds_total,	Container {{

Alert	Severity	Description	Based on	Threshold
than 80%		for pod {{ $labels.pod }}.	container_spec_cpu_period	$labels.container }} CPU usage exceeded 80% for 5 minutes.
Pod memory greater than 65%	Warning	High memory usage for pod {{ $labels.pod }}.	container_memory_working_set_bytes, kube_pod_container_resource_requests_memory_bytes	Container {{ $labels.container }} memory usage exceeded 65% for 5 minutes.
Pod memory greater than 80%	Critical	Critical memory usage for pod {{ $labels.pod }}.	container_memory_working_set_bytes, kube_pod_container_resource_requests_memory_bytes	Container {{ $labels.container }} memory usage exceeded 80% for 5 minutes.
Too many Kafka pending events	Critical	Actions: • Ensure there are no issues with Kafka or {{ $labels.pod }} pod's CPU and network.	kafka_producer_queue_depth	Too many Kafka producer pending events for pod {{ $labels.pod }} (more than 100 in 5 minutes).