## CLAIMS

1. A computer-implemented method, comprising operations for:

receiving a description of a multi-tiered edge topology with a plurality of nodes, a multimedia stream analytics composition that is executed on the plurality of nodes, and performance objectives, wherein the multimedia stream analytics composition includes tasks that use analytic models, wherein the analytic models comprise neural networks;

optimizing at least one of the analytic models to form a list of optimized analytic models by:

for each unique pair of an optimization technique of a plurality of optimization techniques and a precision of a parameter of that analytic model of a plurality of precisions,

generating an optimized analytic model based on the optimization technique and the precision; and

adding the optimized analytic model to the list;

benchmark testing the optimized analytic models on the list using test data and the plurality of nodes in the multi-tiered edge topology;

clustering the optimized analytic models on the list based on benchmark results to form clusters of optimized analytic models;

selecting a representative optimized analytic model from each of the clusters of optimized analytic models;

determining a configuration recommendation that indicates deployment of the tasks and of each selected representative optimized analytic model on the plurality of nodes to meet the performance objectives;

generating one or more workflows from the configuration recommendation to execute the tasks of the multimedia stream analytics composition on nodes of different tiers of the multi-tiered edge topology; and

executing the one or more workflows on the plurality of nodes to generate output for the multimedia stream analytics composition.

2. The computer-implemented method of claim 1, wherein the performance objectives comprise any combination of accuracy, precision, recall, and throughput.

3. The computer-implemented method of claim 1, wherein optimizing the analytic models comprises compressing the analytic models.

4. The computer-implemented method of claim 1, wherein one or more of images and metadata are transmitted between the nodes of the plurality of nodes at different transmission speeds.

5. The computer-implemented method of claim 1, further comprising operations for:

for each node in the multi-tiered edge topology, defining an objective function based on performance objectives for performance, accuracy, precision, and recall;

for each node, defining a cumulative objective function based on a cumulative performance, a cumulative accuracy, a cumulative precision, and a cumulative recall; and

creating an overall objective function for the multi-tiered edge topology based on the cumulative objective function of each node.

6. The computer-implemented method of claim 1, wherein the plurality of nodes includes a central server in a cloud infrastructure, wherein the central server stores one of images and metadata for the multimedia stream analytics composition in a data store.

7. The computer-implemented method of claim 1, wherein a Software as a Service (SaaS) is configured to perform the operations of the method.

8. A computer program product, the computer program product comprising a computer readable storage medium having program code embodied therewith, the program code executable by at least one processor to perform operations for:

receiving a description of a multi-tiered edge topology with a plurality of nodes, a multimedia stream analytics composition that is executed on the plurality of nodes, and performance objectives, wherein the multimedia stream analytics composition includes tasks that use analytic models, wherein the analytic models comprise neural networks;

optimizing at least one of the analytic models to form a list of optimized analytic models by:

for each unique pair of an optimization technique of a plurality of optimization techniques and a precision of a parameter of that analytic model of a plurality of precisions,

generating an optimized analytic model based on the optimization technique and the precision; and

adding the optimized analytic model to the list;

benchmark testing the optimized analytic models on the list using test data and the plurality of nodes in the multi-tiered edge topology;

clustering the optimized analytic models on the list based on benchmark results o form clusters of optimized analytic models;

selecting a representative optimized analytic model from each of the clusters of optimized analytic models;

determining a configuration recommendation that indicates deployment of the tasks and of each selected representative optimized analytic model on the plurality of nodes to meet the performance objectives;

generating one or more workflows from the configuration recommendation to execute the tasks of the multimedia stream analytics composition on nodes of different tiers of the multi-tiered edge topology; and

executing the one or more workflows on the plurality of nodes to generate output for the multimedia stream analytics composition.

9. The computer program product of claim 8, wherein the performance objectives comprise any combination of accuracy, precision, recall, and throughput.

10. The computer program product of claim 8, wherein optimizing the analytic models comprises compressing the analytic models.

11. The computer program product of claim 8, wherein one or more of images and metadata are transmitted between the nodes of the plurality of nodes at different transmission speeds.

12. The computer program product of claim 8, wherein the program code is executable by the at least one processor to perform further operations for:

for each node in the multi-tiered edge topology, defining an objective function based on performance objectives for performance, accuracy, precision, and recall;

for each node, defining a cumulative objective function based on a cumulative performance, a cumulative accuracy, a cumulative precision, and a cumulative recall; and

creating an overall objective function for the multi-tiered edge topology based on the cumulative objective function of each node.

13. The computer program product of claim 8, wherein the plurality of nodes includes a central server in a cloud infrastructure, wherein the central server stores one of images and metadata for the multimedia stream analytics composition in a data store.

14. The computer program product of claim 8, wherein a Software as a Service (SaaS) is configured to perform the operations of the computer program product.

15. A computer system, comprising:

one or more processors, one or more computer-readable memories and one or more computer-readable, tangible storage devices; and

program instructions, stored on at least one of the one or more computer-readable, tangible storage devices for execution by at least one of the one or more processors via at least one of the one or more memories, to perform operations comprising:

receiving a description of a multi-tiered edge topology with a plurality of nodes, a multimedia stream analytics composition that is executed on the plurality of nodes, and performance objectives, wherein the multimedia stream analytics composition includes tasks that use analytic models, wherein the analytic models comprise neural networks;

optimizing at least one of the analytic models to form a list of optimized analytic models by:

for each unique pair of an optimization technique of a plurality of optimization techniques and a precision of a parameter of that analytic model of a plurality of precisions,

generating an optimized analytic model based on the optimization technique and the precision; and

adding the optimized analytic model to the list;

benchmark testing the optimized analytic models on the list using test data and the plurality of nodes in the multi-tiered edge topology;

clustering the optimized analytic models on the list based on benchmark results to form clusters of optimized analytic models;

selecting a representative optimized analytic model from each of the clusters of optimized analytic models;

determining a configuration recommendation that indicates deployment of the tasks and of each selected representative optimized analytic model on the plurality of nodes to meet the performance objectives;

generating one or more workflows from the configuration recommendation to execute the tasks of the multimedia stream analytics composition on nodes of different tiers of the multi-tiered edge topology; and

executing the one or more workflows on the plurality of nodes to generate output for the multimedia stream analytics composition.

16. The computer system of claim 15, wherein the performance objectives comprise any combination of accuracy, precision, recall, and throughput.

17. The computer system of claim 15, wherein optimizing the analytic models comprises compressing the analytic models.

18. The computer system of claim 15, wherein one or more of images and metadata are transmitted between the nodes of the plurality of nodes at different transmission speeds.

19. The computer system of claim 15, wherein the operations further comprise:

for each node in the multi-tiered edge topology, defining an objective function based on performance objectives for performance, accuracy, precision, and recall;

for each node, defining a cumulative objective function based on a cumulative performance, a cumulative accuracy, a cumulative precision, and a cumulative recall; and

creating an overall objective function for the multi-tiered edge topology based on the cumulative objective function of each node.

20. The computer system of claim 15, wherein the plurality of nodes includes a central server in a cloud infrastructure, wherein the central server stores one of images and metadata for the multimedia stream analytics composition in a data store.

21. The computer system of claim 15, wherein a Software as a Service (SaaS) is configured to perform the operations of the computer system.

22. The computer-implemented method of claim 1, further comprising operations for:

performing fine-tuning by trying different combinations of representative optimized analytic models from each of the clusters of optimized analytic models.

23. The computer program product of claim 8, wherein the program code is executable by the at least one processor to perform further operations for:

performing fine-tuning by trying different combinations of representative optimized analytic models from each of the clusters of optimized analytic models.

24. The computer system of claim 15, wherein the operations further comprise:

performing fine-tuning by trying different combinations of representative optimized analytic models from each of the clusters of optimized analytic models.