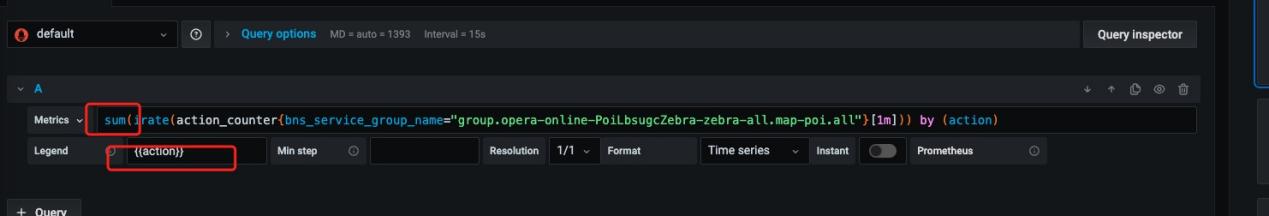
# 普罗米修斯

http\_server\_requests\_count{api="/llmslaver/prompt",err="0",method="POST",statusCode="200"} 4

http\_server\_requests\_count{api="/llmslaver/prompt",err="10013",method="POST",statusCode="200"} 3

可以看到 上面两个点的err不一样,就分出来了两条记录, 所有语句里面如果只对llmslaver/prompt统计的话要加个sum



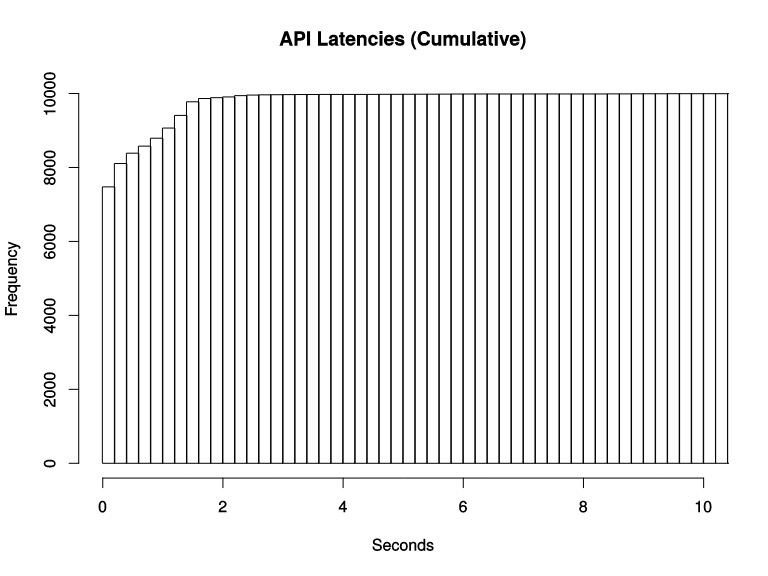
上面这个sum是吧多个实例的加起来,下面的action表示的是横坐标.

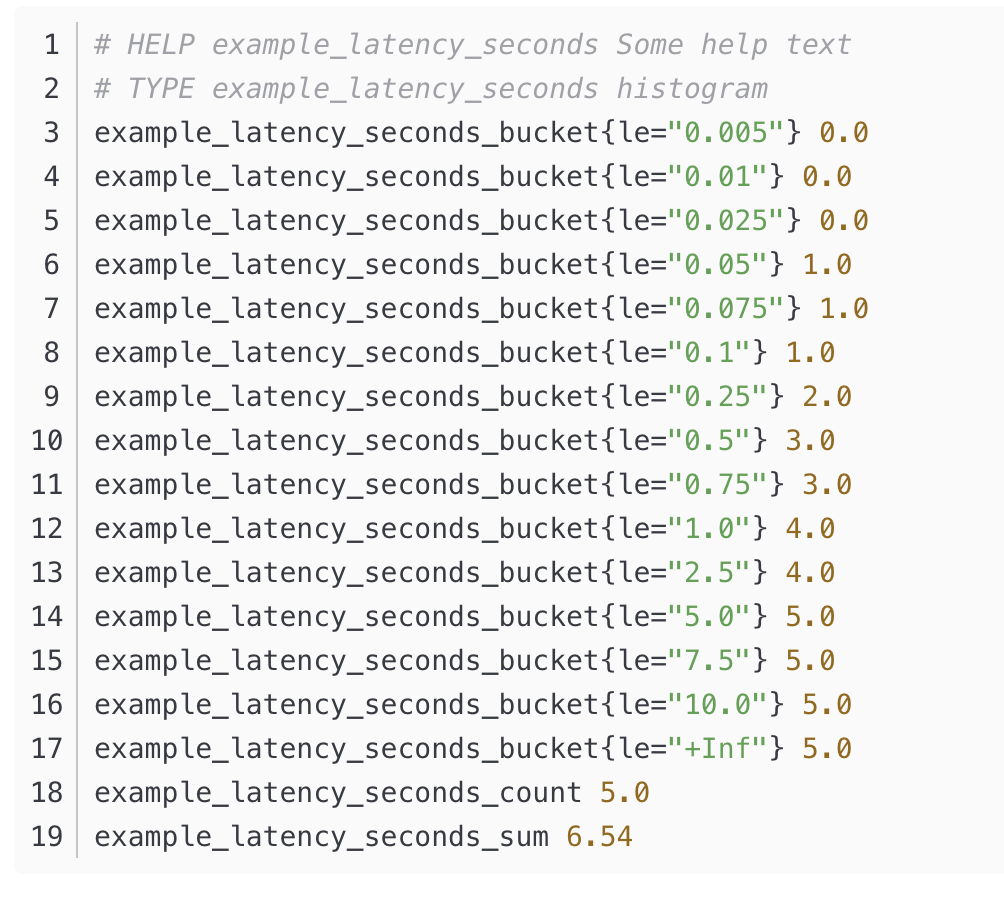
与Summary类型的指标相似之处在于Histogram类型的样本同样会反应当前指标的记录的总数(以\_count作为后缀)以及其值的总量（以\_sum作为后缀）。不同在于Histogram指标直接反应了在不同区间内样本的个数，区间通过标签len进行定义。

同时对于Histogram的指标，我们还可以通过histogram\_quantile()函数计算出其值的分位数。不同在于Histogram通过histogram\_quantile函数是在服务器端计算的分位数。 而Sumamry的分位数则是直接在客户端计算完成。因此对于分位数的计算而言，Summary在通过PromQL进行查询时有更好的性能表现，而Histogram则会消耗更多的资源。反之对于客户端而言Histogram消耗的资源更少。在选择这两种方式时用户应该按照自己的实际场景进行选择。

Histogram

是累计直方图





累计直方图就是后面的会把前面的都加上.

p95就是看长尾的

histogram\_quantile(0.95, sum(rate(request\_cost\_bucket{bns\_service\_group\_name="group.opera-online-PoiTaojinArch-TaojinArch-all.map-poi.all"}[3m])) by (le,api))

平均耗时

avg(rate(action\_cost\_sum{bns\_service\_group\_name="group.opera-online-PoiOperatorMarket-OperatorMarket-all.map-poi.all",group="$group"}[5m]) / rate(action\_cost\_count{bns\_service\_group\_name="group.opera-online-PoiOperatorMarket-OperatorMarket-all.map-poi.all",group="$group"}[5m])) by (action)