在线推理增加模型管理服务和分布式配置功能

在实际生产中使用Fate框架时，往往会涉及到很多团队之间的合作，就会遇到各种问题。下面先总结一下使用fate\_serving遇到的问题。

遇到的问题

1. 训练环境和在线推理（生产）环境网络不通，无法将模型导入到推理环境。
2. 发布或者绑定模型在server.conf配置文件中配置所有的serving-server的节点，因为训练环境和推理可能处于不同的网络环境中，这其中涉及交互的机器越多，出现错误的可能性也就越大。
3. 推理服务扩容时还需进行load和bind 操作。

需要功能

1. 下线（禁止调用）某个已上线的模型。
2. 好多配置其实是冗余的，应该有分布式配置的功能。

Model\_manager功能

1. server节点启动时，主动去model\_manager拉取所有已经上线的模型。

**应用场景：**扩容就不需要再进行load和bind 操作。

1. 模型更新后（fate\_flow发布新模型到推理服务、下线某个模型等），主动通知server节点变更模型状态。

**应用场景**：实现下线（禁止调用）某个已上线的模型等。

1. model\_manager可以从本地加载模型文件。

**应用场景：**训练环境和推理环境网络不通问题。

1. fate\_flow只与model\_manager交互，不同的网络环境之间交互的次数及机器越少，出现的错误的可能性也就越少。

config功能

1. 分布式配置功能。

设计图

